


3 1761 10374381 1



Digitized by the Internet Archive
in 2023 with funding from
University of Toronto

<https://archive.org/details/31761103743811>

12-001



Government
Publications

SURVEY METHODOLOGY

Catalogue 12-001

A JOURNAL
PUBLISHED BY
STATISTICS CANADA

DECEMBER 1993

•

VOLUME 19

•

NUMBER 2



Statistics
Canada

Statistique
Canada

Canada



SURVEY METHODOLOGY

A JOURNAL
PUBLISHED BY
STATISTICS CANADA

DECEMBER 1993 • VOLUME 19 • NUMBER 2

Published by authority of the Minister
responsible for Statistics Canada

© Minister of Industry,
Science and Technology, 1993

All rights reserved. No part of this publication may be reproduced,
stored in a retrieval system or transmitted in any form or by any
means, electronic, mechanical, photocopying, recording or otherwise
without prior written permission from Licence Services,
Marketing Division, Statistics Canada,
Ottawa, Ontario, Canada K1A 0T6.

December 1993

Price: Canada: \$45.00
United States: US\$50.00
Other Countries: US\$55.00

Catalogue No. 12-001

ISSN 0714-0045

Ottawa



Statistics Canada
Statistique Canada

Canada

SURVEY METHODOLOGY

A Journal of Statistics Canada

The Survey Methodology Journal is abstracted in The Survey Statistician and Statistical Theory and Methods Abstracts and is referenced in the Current Index to Statistics, and Journal Contents in Qualitative Methods.

MANAGEMENT BOARD

Chairman G.J. Brackstone

Members B.N. Chinnappa C. Patrick
G.J.C. Hole D. Roy
F. Mayda (Production Manager) M.P. Singh
R. Platek (Past Chairman)

EDITORIAL BOARD

Editor M.P. Singh, *Statistics Canada*

Associate Editors

D.R. Bellhouse, *University of Western Ontario*

D. Binder, *Statistics Canada*

M. Colledge, *Statistics Canada*

E.B. Dagum, *Statistics Canada*

J.-C. Deville, *INSEE*

D. Drew, *Statistics Canada*

R.E. Fay, *U.S. Bureau of the Census*

W.A. Fuller, *Iowa State University*

J.F. Gentleman, *Statistics Canada*

M. Gonzalez, *U.S. Office of Management and Budget*

R.M. Groves, *U.S. Bureau of the Census*

D. Holt, *University of Southampton*

G. Kalton, *University of Michigan*

D. Pfeffermann, *Hebrew University*

J.N.K. Rao, *Carleton University*

L.-P. Rivest, *Université Laval*

D.B. Rubin, *Harvard University*

I. Sande, *Bell Communications Research, U.S.A.*

C.-E. Särndal, *Université de Montréal*

W.L. Schaible, *U.S. Bureau of Labor Statistics*

F.J. Scheuren, *U.S. Internal Revenue Service*

J. Sedransk, *State University of New York*

C.M. Suchindran, *University of North Carolina*

J. Waksberg, *Westat Inc.*

K.M. Wolter, *A.C. Nielsen, U.S.A.*

Assistant Editors

N. Laniel, P. Lavallée, L. Mach and H. Mantel, *Statistics Canada*

EDITORIAL POLICY

The Survey Methodology Journal publishes articles dealing with various aspects of statistical development relevant to a statistical agency, such as design issues in the context of practical constraints, use of different data sources and collection techniques, total survey error, survey evaluation, research in survey methodology, time series analysis, seasonal adjustment, demographic studies, data integration, estimation and data analysis methods, and general survey systems development. The emphasis is placed on the development and evaluation of specific methodologies as applied to data collection or the data themselves. All papers will be refereed. However, the authors retain full responsibility for the contents of their papers and opinions expressed are not necessarily those of the Editorial Board or of Statistics Canada.

Submission of Manuscripts

The Survey Methodology Journal is published twice a year. Authors are invited to submit their manuscripts in either of the two official languages, English or French to the Editor, Dr. M.P. Singh, Social Survey Methods Division, Statistics Canada, Tunney's Pasture, Ottawa, Ontario, Canada K1A 0T6. Four nonreturnable copies of each manuscript prepared following the guidelines given in the Journal are requested.

Subscription Rates

The price of the Survey Methodology Journal (Catalogue No. 12-001) is \$45 per year in Canada, US \$50 in the United States, and US \$55 per year for other countries. Subscription order should be sent to Publication Sales, Statistics Canada, Ottawa, Ontario, Canada K1A 0T6. A reduced price is available to members of the American Statistical Association, the International Association of Survey Statisticians, and the Statistical Society of Canada.

SURVEY METHODOLOGY

A Journal of Statistics Canada

Volume 19, Number 2, December 1993

CONTENTS

In This Issue	125
 P.P. BIEMER and D. ATKINSON Estimation of Measurement Bias Using a Model Prediction Approach	127
 J.B. ARMSTRONG and J.E. MAYDA Model-Based Estimation of Record Linkage Error Rates	137
 D. PFEFFERMANN and S.R. BLEUER Robust Joint Modelling of Labour Force Series of Small Areas	149
 I.U.H. MIAN and N. LANIEL Maximum Likelihood Estimation of Constant Multiplicative Bias Benchmarking Model with Application ..	165
 J.-C. DEVILLE Optimum Two-Stage Sample Design for Ratio Estimators: Application to Quality Control – 1990 French Census	173
 R.J. CASADY and R. VALLIANT Conditional Properties of Post-Stratified Estimators Under Normal Theory	183
 S. BANDYOPADHYAY and A.K. ADHIKARI Sampling from Imperfect Frames with Unknown Amount Of Duplication	193
 F.A. ROESCH, JR., E.J. GREEN and C.T. SCOTT An Alternative View of Forest Sampling	199
 G. KALTON and C.F. CITRO Panel Surveys: Adding the Fourth Dimension	205
 Acknowledgements	217

In This Issue

Papers covering a variety of topics are included in this issue of *Survey Methodology*. In the first paper, Biemer and Atkinson present a general methodology for constructing and evaluating model prediction estimators of measurement bias for a stratified two-phase design with simple random sampling in each phase. For evaluation, they extended the bootstrap methodology of Bickel and Freedman to two-phase sampling. The example used for illustration indicates that improvements over the traditional net difference estimator and thus savings in the cost of reinterview surveys are possible.

The paper by Armstrong and Mayda was originally intended for the special section *Record Linkage and Statistical Matching*. The authors consider model based estimation of classification error rates in record linkage. The class of models considered allows for non-independence of match status of different matching fields within a record pair. Estimation methods are developed and different methods of error rate estimation are compared using both synthetic and real data.

Pfeffermann and Bleuer consider estimation for small areas using data from a rotating panel survey over time. Their approach is model based, with a state space model for the population values over time and separate autoregressive models for the survey error series from each panel. To achieve a measure of robustness, the small area estimators are further constrained to add up to direct survey estimators within pre-defined larger areas. The approach is demonstrated using Canadian Labour Force Survey data for the Atlantic provinces.

Mian and Laniel discuss two iterative procedures to find the maximum likelihood estimates of a non-linear benchmarking model that seems suitable for economic time series from large sample surveys. Closed form expressions for the asymptotic variances and covariances of the benchmarked series and of the fitted values are also provided. The methodology is illustrated using Canadian retail trade data.

Deville uses superpopulation models to anticipate, before data collection, the variances of estimates of ratios. Based on models that are both simple and realistic, he produces expressions of varying complexity and then optimizes them. He deals with the problem of estimating the frequency of errors in the population of forms collected during the quality control of the French census.

Asymptotic techniques are used by Casady and Valliant to study post-stratification from a design-based, conditional point of view. The authors derive the large sample bias and mean squared error of the standard post-stratified estimator, the Horvitz-Thompson estimator, a ratio estimator and a new post-stratified regression estimator. The developed theory is empirically tested using real and artificial populations. The problem of bias due to defective frames is also addressed.

Bandyopadhyay and Adhikari study estimation based on frames where some units are listed more than once, each time with a different identification. The mean square errors of estimators from imperfect and perfect frames are compared. Estimation of a population ratio, mean and total when no auxiliary information is available on the frame is considered.

Roesch, Green and Scott present a generalized concept for all of the commonly used methods of forest sampling. The concept views the forest as a two-dimensional picture which is cut up into pieces like a jigsaw puzzle, with the pieces defined by the individual selection probabilities of the trees in the forest.

The paper by Kalton and Citro is a revised version of the keynote address given at the Statistics Canada Symposium 92 on longitudinal surveys. The paper discusses how different designs for surveys over time satisfy various analytic objectives. The author then concentrates on panel surveys and talks about decisions that need to be made when designing them.

Estimation of Measurement Bias Using a Model Prediction Approach

PAUL P. BIEMER and DALE ATKINSON¹

ABSTRACT

Methods for estimating response bias in surveys require “unbiased” remeasurements for at least a subsample of observations. The usual estimator of response bias is the difference between the mean of the original observations and the mean of the unbiased observations. In this article, we explore a number of alternative estimators of response bias derived from a model prediction approach. The assumed sampling design is a stratified two-phase design implementing simple random sampling in each phase. We assume that the characteristic, y , is observed for each unit selected in phase 1 while the true value of the characteristic, μ , is obtained for each unit in the subsample selected at phase 2. We further assume that an auxiliary variable x is known for each unit in the phase 1 sample and that the population total of x is known. A number of models relating y , μ and x are assumed which yield alternative estimators of $E(y - \mu)$, the response bias. The estimators are evaluated using a bootstrap procedure for estimating variance, bias, and mean squared error. Our bootstrap procedure is an extension of the Bickel-Freedman single phase method to the case of a stratified two-phase design. As an illustration, the methodology is applied to data from the National Agricultural Statistics Service reinterview program. For these data, we show that the usual difference estimator is outperformed by the model-assisted estimator suggested by Särndal, Swensson and Wretman (1991), thus indicating that improvements over the traditional estimator are possible using the model prediction approach.

KEY WORDS: Reinterview; Repeated measures; Response error; Bootstrap.

1. INTRODUCTION

It is well-known in the survey literature that when responses are obtained from respondents in sample surveys, the observed values of measured characteristics may differ markedly from the true values of the characteristics. Evidence of these so-called measurement errors in surveys has been collected in a number of ways. For example, the recorded response may be checked for accuracy against administrative records or legal documents within which the true (or at least a more accurate) value of the characteristic is contained. An alternative approach relies on revised reports from respondents via reinterviews. In a reinterview, a respondent is recontacted for the purpose of conducting a second interview regarding the same characteristics measured in the first interview. Rather than simply repeating the original questions in the interview, there may be extensive probes designed to elicit more accurate responses, or the respondent may be instructed to consult written records for the “book values” of the characteristics. For some reinterview surveys, discrepancies between the first and second interviews are reconciled with the respondent until the interviewer is satisfied that a correct answer has been obtained. Forsman and Schreiner (1991) provide an overview of the literature for these types of reinterviews. Other means of checking the accuracy of survey responses include: (a) comparing the survey

statistics (*i.e.*, means, totals, proportions, *etc.*) to statistics from external sources that are more accurate; (b) using experimental designs to estimate the effects on survey estimates of interviewers and other survey personnel; and (c) checking the results within the same survey for internal consistency.

The focus of the current work is on estimators of measurement bias from data collected in true value remeasurement studies, *i.e.*, record check and reinterview studies, where the objective is to obtain the true value of the characteristic at, perhaps, a much greater cost per measurement than that of the original observation.

Because of the high costs typically involved in conducting reinterview studies, repeated measurements are usually obtained for only a small fraction of the original survey sample. While the sample size may be quite adequate for estimating biases at the national and regional levels, they may not be adequate for estimating the error associated with small subpopulations or rare survey characteristics. In this paper, our objective is to consider estimators of response bias having better mean squared error properties than the traditional estimators. The basic idea behind our approach can be described as follows.

In a typical remeasurement study, a random subsample of the survey respondents is selected and, through some means, the true values of the characteristics of interest are ascertained. Let n_1 denote the number of respondents to

¹ Paul P. Biemer, Principal Scientist, Center for Survey Research, Research Triangle Institute, Research Triangle Park, NC 27709; Dale Atkinson, Supervisory Mathematical Statistician, National Agricultural Statistics Service, 3251 Old Lee Highway, Room 305, Fairfax, Va 22030.

the first survey and let n_2 denote the number selected for the subsample or evaluation sample. The usual estimator of response bias is the net difference rate, computed for the n_2 respondents in the evaluation sample as

$$\text{NDR} = \bar{y}_2 - \bar{\mu}_2, \quad (1.1)$$

where \bar{y}_2 is the sample mean of original responses and $\bar{\mu}_2$ is the sample mean of the true measurements. A disadvantage of the NDR is that it excludes information on the $n_1 - n_2$ units in the original survey who were not included in the remeasurement study. Further, the estimator does not incorporate information on auxiliary variables, x , which may be combined with the information on y and μ available from the survey to provide a more precise estimator of response bias.

Given that we have a stratified, two-phase sample design and resulting data (y, μ, x) , our objective is to determine the “best” estimator of measurement bias given these data. Our basic approach is to identify a model for the true value, μ_i , which is a function of the observed values, $y_i, i = 1, \dots, n_1$, and any auxiliary information, x , that may be available for the population. The model is then used to predict μ_i for all units in the population for which μ_i is unknown. These predictions can then be used to obtain estimates of the true population mean, total, or proportion. Thus, estimators of the response bias for these parameters can be derived from the main survey. Since the approach provides a prediction equation for μ_i which is a function of the observations, estimators of response bias can be computed for areas having small sample sizes. In this case, the prediction equation for μ_i may be augmented by other respondent variables such as demographic characteristics, type of unit, unit size, geographic characteristics, and so on.

The basic estimation and evaluation theory for a prediction approach to the estimation of response bias is presented in the following sections. Under stratified random sampling, estimators of means and totals, their variances and their mean squared errors are provided. Results from application to National Agricultural Statistics Service (NASS) data are also presented.

2. METHODOLOGY FOR ESTIMATION AND EVALUATION

2.1 The Measurement Error Model

To fix the ideas, we shall consider the case of simple random sampling without replacement (SRSWOR) from a single population. Generalizations to stratified random sampling are straightforward and will be considered subsequently.

Let $U = \{1, 2, \dots, N\}$ denote the label set for the population and let $S_1 = \{1, 2, \dots, n_1\}$, without loss of generality, denote the label set for the first phase SRSWOR sample of n_1 units from U .

For $y_i, i \in S_1$, assume the model

$$y_i = \gamma_0 + \gamma\mu_i + \epsilon_i, \quad (2.1)$$

where μ_i is the true value of the measured characteristic, γ_0 and γ are constants, and ϵ_i is an independent error term having zero expectation and conditional variance, $\sigma_{\epsilon_i}^2$.

Since the focus of our investigation is on the bias associated with the measurements y_i , consider the expectation of y_i . Let $E(y_i | i)$ denote the conditional expectation of y_i over the distribution of the ϵ_i holding the unit i fixed and let $E(y_i) = E_i[E(y_i | i)]$ denote the expectation of $E(y_i | i)$ over the sampling distribution. Then, for a given unit, i ,

$$E(y_i | i) = \gamma_0 + \gamma\mu_i \quad (2.2)$$

and, hence, the unconditional expectation is

$$E(y_i) = \gamma_0 + \gamma\bar{M}, \quad (2.3)$$

where $\bar{M} = \sum_{i=1}^N \mu_i / N$. Thus, the measurement bias is

$$\bar{B} = E(y_i - \mu_i) = \gamma_0 + (\gamma - 1)\bar{M}. \quad (2.4)$$

The parameter, γ_0 , is a constant bias term that does not depend upon the magnitude of \bar{M} . Note that this definition of γ_0 is consistent with the usual definition of measurement bias obtained from the simple model

$$y_i = \mu_i + \epsilon_i, \quad (2.5)$$

with $\epsilon_i \sim (\gamma_0, \sigma_{\epsilon_i}^2)$. (See, for example, Biemer and Stokes 1991.)

Consider the estimation of \bar{B} . Assume that a subsample of size n_2 of the original n_1 sample units is selected and the true value, μ_i , is measured for these n_2 units. The true value may be ascertained either by a reinterview, a record check, interviewer observation, or some other means. Let $S_2 \subseteq S_1$ denote this so-called second phase sample. The usual estimator of the measurement bias is the NDR defined in (1.1). If the assumption that “the true value, μ_i , is observed in phase 2, for all $i \in S_2$ ” is satisfied, then the NDR is an unbiased estimator of \bar{B} . It may further be shown that the variance of the NDR is

$$E \left\{ \left(1 - \frac{n_2}{n_1} \right) \frac{s_{\mu}^2}{n_2} \left(1 - \frac{s_{\mu y}^2}{s_y^2 s_{\mu}^2} \right) + \left(1 - \frac{n_2}{n_1} \right) \frac{s_y^2}{n_2} (1 - r)^2 \right\}, \quad (2.6)$$

where $s_\mu^2 = \sum_{j \in S_2} (\mu_j - \bar{\mu}_2)^2 / (n_2 - 1)$ with analogous definitions for s_y^2 and $s_{\mu y}$, and $r = s_{\mu y} / s_y^2$.

The NDR may be suboptimal in a number of situations which occur with some frequency. To see this, consider estimators of \bar{B} of the form

$$\bar{b}_{ga} = \bar{y}_g - \bar{\mu}_{Ra}, \quad (2.7)$$

where $\bar{y}_g = \sum_{j \in S_g} y_j / n_g$, $g = 1, 2$,

$$\bar{\mu}_{Ra} = \bar{\mu}_2 + a(\bar{y}_1 - \bar{y}_2) \quad (2.8)$$

and $\bar{\mu}_2 = \sum_{j \in S_2} \mu_j / n_2$, for a a constant given the subsample, S_1 . It can be shown that the value of a that minimizes $\text{Var}(\bar{b}_{ga})$ is

$$\begin{aligned} a &= r & \text{for } g &= 1, \\ \text{or } a &= r - 1 & \text{for } g &= 2. \end{aligned} \quad (2.9)$$

Thus, for $g = 1$ or 2 , the “optimal” choice of \bar{b}_{ga} is

$$\bar{b}_{\text{opt}} = \bar{y}_1 - [\bar{\mu}_2 + r(\bar{y}_1 - \bar{y}_2)], \quad (2.10)$$

which differs from the NDR by the term $(r - 1)(\bar{y}_1 - \bar{y}_2)$. Since, in general, $\bar{y}_1 \neq \bar{y}_2$, NDR is optimal only if $r = 1$. It can be shown that this corresponds to the case where γ_1 in (2.1) is 1.

In this paper we shall explore alternatives to the NDR which incorporate information on y for units in the set $S_1 \sim S_2$ as well as information on some auxiliary variable, x . To illustrate the concepts, we shall restrict ourselves to “no-intercept” linear models initially, *i.e.*, models for which $\gamma_0 = 0$ in (2.1). This important class of models includes the difference estimator as well as ratio estimators.

2.2 Model Prediction Approaches To Estimation

Model prediction approaches to the estimation of population parameters in finite population sampling are well-documented in the literature. Cochran (1977) and other authors have demonstrated the model-based foundations of the ubiquitous ratio estimator. There is also considerable literature on the choice between using weights that are derived from explicit model assumptions in estimation for complex surveys or eliminating the sample weights. Proponents of so-called model-based estimation recommend against the use of weights in parameter estimation (see, for example, Royall and Herson 1973; and Royall and Cumberland 1981). They contend that the probabilities of selection in finite population sampling, whether equal or unequal, are irrelevant once the sample is produced. The reliability criteria used by model-based samples are derived from the model distributional assumptions rather than sampling distributions. If an appropriate

model is chosen to describe the relationship between the response variable and other measured survey variables, “model-unbiased” estimators of the population parameters may be obtained which have greater reliability than estimators which incorporate weights.

On the other side of the controversy are the design-based samplers. Instead of the model-based assumptions, design-based samplers assume that an estimator from a survey is a single realization from a large population of potential realizations of the estimator, where each potential realization depends upon the selected sample. The distribution of the values of the estimator when all possible samples that may be selected by the sampling scheme are considered is referred to as the sampling distribution of the estimator. Criteria for evaluating estimators under the design-based approach then consider the properties of the sampling distributions of the estimators. Under this approach, weighting of the estimators is required to achieve unbiasedness if unequal probability sampling is used.

Although the estimators of \bar{B} considered here represent all three classes of estimators, the objective of this paper is not necessarily to compare design-based, model-assisted, and model-based estimators. Rather, we first seek to develop a systematic approach for evaluating alternative estimators for a given two-phase sample design. The problem considered is the following: Given a two-phase sample design and estimators of $B = N\bar{B}$ denoted by $\hat{B}_1, \hat{B}_2, \dots, \hat{B}_p$, how does an analyst identify which estimator minimizes the mean squared error? A second objective of the article is to specify a number of alternative estimators, and apply a systematic approach for evaluating the estimators. As an illustration, the methodology will be applied to data from the National Agricultural Statistics Service’s December 1990 Agricultural Survey.

2.3 The Estimators Considered in Our Study

Extending the previously developed notation to stratified, two-phase designs, let N_h denote the size of the h th stratum, for $h = 1, \dots, L$. A two-phase sample is selected in each stratum using simple random sampling at each phase. Let n_{1h} and $n_{2h} \leq n_{1h}$ denote the phase 1 and phase 2 sample sizes, respectively, in stratum h . Let S_{1h} and $S_{2h} \subseteq S_{1h}$ denote the label sets for the phase 1 and phase 2 samples, respectively, in stratum h . Assume the following data are either observed or otherwise known:

outcome variables: $y_i \quad \forall i \in S_{1h}$

true values: $\mu_i \quad \forall i \in S_{2h}$

auxiliary variables: $x_i \quad \forall i \in S_{1h}$.

Further assume that $X_h = \sum_{i \in U_h} x_i$ is known for $h = 1, \dots, L$ where U_h is the label set for the h th stratum.

2.3.1 Weighted Estimators of M and B

As a matter of convenience, we shall consider the estimation of the bias for an estimator of a population total denoted by M . The usual estimator of $M = N\bar{M}$ is the unbiased stratified estimator given by

$$\hat{M}_{2st} = \sum_h N_h \bar{\mu}_{2h}, \quad (2.11)$$

where $\bar{\mu}_{2h} = \sum_{i \in S_{2h}} \mu_i / n_{2h}$. The corresponding estimator of $B = N\bar{B}$ is N times the NDR defined in (1.1). For stratified samples, it is

$$\hat{B}_{2st} = \hat{Y}_{2st} - \hat{M}_{2st}, \quad (2.12)$$

where $\hat{Y}_{2st} = \sum_h N_h \bar{y}_{2h}$ and $\bar{y}_{2h} = \sum_{i \in S_{2h}} y_i / n_{2h}$. Note that (2.12) does not incorporate the information on y for units with labels $i \in S_{1h} \sim S_{2h}$. An alternative estimator that uses all the data on y is

$$\hat{B}_{12st} = \hat{Y}_{1st} - \hat{M}_{2st}, \quad (2.13)$$

where $\hat{Y}_{1st} = \sum_h N_h \bar{y}_{1h}$ and $\bar{y}_{1h} = \sum_{i \in S_{1h}} y_i / n_{1h}$.

A number of model-assisted estimators can be specified for two-phase stratified designs. These may take the form of either separate or combined estimators (see, for example, Cochran 1977, pp. 327-330). Further, the ratio adjustments may be applied to either phase 1 or phase 2 stratum-level estimators. Because stratum sample sizes are typically small in two-phase samples, only combined estimators shall be considered here.

As the emphasis in this paper is on the development of the methodology for model-based estimates of measurement bias and their evaluation, we shall consider a simple, special case of the model (2.1); viz., $\gamma_0 = 0$ or the no-intercept model. However, generalizations of the no-intercept methodology to multivariate intercept models do not afford any difficulties and will be considered in a subsequent paper. Thus, letting $\gamma_0 = 0$ in (2.1) we have

$$y_i = \gamma \mu_i + \epsilon_i, \quad (2.14)$$

where γ is an unknown constant and we assume $\epsilon_i \sim (0, \sigma_\epsilon^2 \mu_i)$. The least squares estimator of γ is $\hat{\gamma} = \bar{y}_{2st} / \bar{\mu}_{2st}$, where $\bar{y}_{2st} = \hat{Y}_{2st} / N$ and $\bar{\mu}_{2st} = \hat{M}_{2st} / N$. Thus, a model-assisted estimator of μ_i is $y_i / \hat{\gamma} = \bar{\mu}_{2st} y_i / \bar{y}_{2st}$ and of M is

$$\hat{M}_{2stR} = \frac{\hat{M}_{2st}}{\hat{Y}_{2st}} \hat{Y}_{1st}. \quad (2.15)$$

Using this estimator of M , two estimators of B corresponding to (2.12) and (2.13) are

$$\hat{B}_{2stR} = \hat{Y}_{2st} - \hat{M}_{2stR} \quad (2.16)$$

and

$$\hat{B}_{12stR} = \hat{Y}_{1st} - \hat{M}_{2stR}. \quad (2.17)$$

A third estimator of B can be obtained via the model

$$y_i = \beta x_i + e_i, \quad (2.18)$$

where β is a constant and $e_i \sim (0, \sigma_e^2 x_i)$. This leads to a ratio estimator of Y ,

$$\hat{Y}_{xstR} = \frac{\bar{y}_{1st}}{\bar{x}_{1st}} X. \quad (2.19)$$

Thus, the corresponding estimator of B is

$$\hat{B}_{x2stR} = \hat{Y}_{xstR} - \hat{M}_{2stR}. \quad (2.20)$$

Finally, Särndal, Swensson and Wretman (1992, p. 360) suggest a general estimator of M in two-phase sampling. Applying their equation 9.7.2 to the model in (2.14) under stratified sampling yields

$$\hat{M}_{SSW} = \hat{M}_{2stR} + \frac{\bar{\mu}_{2st}}{\bar{x}_{2st}} (X - \hat{X}_{1st}). \quad (2.21)$$

Note that this estimator is simply (2.15) with the addition of the unbiased estimator of zero. The resulting estimator may have smaller variance than \hat{M}_{2stR} if this term is negatively correlated with \hat{M}_{2stR} . Likewise, their estimator of Y reduces to \hat{Y}_{xstR} defined in (2.19). Thus the corresponding estimator of B is

$$\hat{B}_{SSW} = \hat{Y}_{xstR} - \hat{M}_{SSW}, \quad (2.22)$$

which is identical to $\hat{B}_{SSW} = B_{x2stR}$ plus the second term of the right hand side of (2.21).

2.3.2 Unweighted Estimators of M and B

Rewrite M as

$$M = \sum_{i \in S_2} \mu_i + \sum_{i \in S_1 \sim S_2} \mu_i + \sum_{i \in U \sim S_1} \mu_i \quad (2.23)$$

$$= M_{(2)} + M_{(1 \sim 2)} + M_{(\sim 1)},$$

say, where $S_g = \bigcup_{h=1}^L S_{gh}$, $g = 1, 2$. The strategy for unweighted, model-based estimation is to replace μ_i in $M_{(1 \sim 2)}$ and $M_{(\sim 1)}$ by a prediction, $\hat{\mu}_i$, obtained from a model.

Using the model in (2.14), an estimator of μ_i is

$$\hat{\mu}_i = y_i / \hat{\gamma},$$

where now $\hat{\gamma} = \bar{y}_2 / \bar{\mu}_2$. Thus, an estimator of $M_{(1 \sim 2)}$ is

$$\hat{M}_{(1 \sim 2)} = \frac{\bar{\mu}_2}{\bar{y}_2} \sum_{i \in S_1 \sim S_2} y_i \quad (2.24)$$

$$= \frac{\bar{\mu}_2}{\bar{y}_2} (n_1 \bar{y}_1 - n_2 \bar{y}_2),$$

where $\bar{y}_g = \sum_{i \in S_g} y_i / n_g$, $\bar{\mu}_2 = \sum_{i \in S_2} \mu_i / n_2$, and $n_g = \sum_h n_{gh}$, for $g = 1, 2$. Further, using the model

$$\mu_i = \delta x_i + \xi_i, \quad (2.25)$$

where δ is a constant and $\xi_i \sim (0, \sigma_\xi^2 x_i)$, we obtain

$$\hat{M}_{(\sim 1)} = \frac{\bar{\mu}_2}{\bar{x}_2} X_{U \sim S_1}, \quad (2.26)$$

where $X_{U \sim S_1} = \sum_{i \in U \sim S_1} X_i$. Thus, a model based estimator of M is

$$\begin{aligned} \hat{M}_M &= M_{(2)} + \hat{M}_{(1 \sim 2)} + \hat{M}_{(\sim 1)} \\ &= \hat{M}_{(1)} + \hat{M}_{(\sim 1)}, \end{aligned} \quad (2.27)$$

where $\hat{M}_{(1)} = n_1 \bar{\mu}_2 \bar{y}_1 / \bar{y}_2$.

Likewise, Y can be rewritten as

$$\begin{aligned} Y &= \sum_{i \in S_1} y_i + \sum_{i \in U \sim S_1} y_i \\ &= Y_{(1)} + Y_{(\sim 1)} \end{aligned} \quad (2.28)$$

and we wish to predict y_i in $Y_{(\sim 1)}$. Using the model in (2.18) a model-based estimator of $Y_{(\sim 1)}$ is

$$\hat{Y}_{(\sim 1)} = \frac{\bar{y}_1}{\bar{x}_1} X_{U \sim S_1}$$

and, thus, an estimator of Y is

$$\hat{Y}_M = Y_{(1)} + \hat{Y}_{(\sim 1)}. \quad (2.29)$$

Thus, B is estimated as

$$\hat{B}_M = \hat{Y}_M - \hat{M}_M. \quad (2.30)$$

Versions of \hat{B}_{2stR} , \hat{B}_{12stR} , \hat{B}_{x2stR} and \hat{B}_M which are more robust to model outliers may also be constructed. The corresponding estimators, denoted by \tilde{B}_{2stR} , \tilde{B}_{12stR} , \tilde{B}_{x2stR} and \tilde{B}_M , respectively, may be formed by eliminating those data points which deviate substantially from the model predictions and computing the model-based or model-assisted estimators using the remaining data. To illustrate, consider the estimator \hat{M}_{2stR} in (2.15). For this estimator, let

$$(n_{2h} - 1) s_{res,h}^2 = \sum_{\mu_{hi} \neq 0} \frac{(y_{hi} - \hat{\gamma} \mu_{hi})^2}{\mu_{hi}}, \quad (2.31)$$

denote the sum of squares of residuals for the model (2.14). Then, in calculating the estimator of γ , only those units in $i \in \tilde{S}_{2h}$ where $\tilde{S}_{2h} = \{i \in S_{2h} : |y_{ih} - \hat{\gamma} \mu_{ih}| \leq 3s_{res,h} \sqrt{\mu_{hi}}\}$ are used. Denoting this estimator of γ as $\tilde{\gamma}$, the estimator of M is $\tilde{M}_{2stR} = \hat{Y}_{1st} / \tilde{\gamma}$ where $\tilde{\gamma} = \tilde{y}_{2st} / \tilde{\mu}_{2st}$ and $\tilde{\mu}_{2st}$ and \tilde{y}_{2st} are the stratified means of μ_i and y_i for $i \in \tilde{S}_{2h}$. The other robust model prediction estimators may be computed analogously.

Many other unweighted, model-based estimators may be explored in the context of our two-phase design. For example, an intercept term may be added to models (2.14), (2.18), and (2.25). Further, slope and intercept parameters may be specified separately for each stratum or combination of strata.

2.4 Estimation of Mean Squared Errors Using Bootstrap Estimators

Although it is possible, under the appropriate design-based or model-based assumptions, to derive closed form analytical estimates of the variance of the estimators we are considering in this study, we have elected instead to use a computer-intensive resampling method. First, we seek a method which is easy to apply since there are potentially many estimators which will be considered in our study. Secondly, it is important to evaluate each estimator using the same criteria and a consistent method of variance estimation is essential to achieving this objective. Thus, it is essential that we employ a variance estimation method which can be applied to estimators of any complexity, under assumptions which are consistent and which do not rely upon any model assumptions. It is well-known that model-based variance estimation approaches are quite sensitive to model failure (see, for example, Royall and Herson 1973; Royall and Cumberland 1978; and Hansen, Madow and Tepping 1983). Royall and Cumberland (1981) discuss several bias relevant alternatives including the jackknife variance estimator.

Our approach is similar to that of Royall and Cumberland except rather than using a jackknife estimator, we employ a bootstrap estimator of the variance. For independent and identically distributed observations, Efron and Gong (1983) show that the bootstrap and the jackknife variance estimators differ by a factor of $n/(n-1)$ for samples of size n . Thus, the robustness properties Royall and Cumberland demonstrate for the jackknife estimator also hold for the bootstrap estimator.

Other properties of the bootstrap estimator have led us to choose it above other resampling methods. The jackknife and balance repeated replication (BRR) methods are not easily modified for the two-phase sampling design of

our study. However, the bootstrap is readily adaptable to two-phase sampling. Further, Rao and Wu (1988) provide evidence from a simulation study that the coverage properties of bootstrap confidence intervals in complex sampling compare favorably to the jackknife and BRR.

Our general approach extends the method developed by Bickel and Freedman (1984) for single phase, stratified sampling, to two-phase stratified sampling. Since the bootstrap procedure is implemented independently for each stratum, we shall, for simplicity, describe the method for the single stratum case.

2.4.1 Estimation of Variance

Extending the bootstrap method to two-phase sampling is not simply a matter of subsampling the single phase bootstrap samples. Recall that true values are known only for the units in S_2 and, therefore, the bootstrap sampling scheme must necessarily confine the selection to units in S_2 . Therefore, let S_1 and S_2 denote the phase 1 and phase 2 samples, respectively, selected from U using SRSWOR. Let $S_{1\sim 2}$ denote the label set, $S_1 \sim S_2$. Let $\hat{\Theta} = \hat{\Theta}(S_{1\sim 2}, S_2)$ denote an estimator of Θ which may be a function of the observations corresponding to units in both S_2 and $S_{1\sim 2}$. Define N , n_1 , n_2 and $n_{1\sim 2}$ as the sizes of sets U , S_1 , S_2 and $S_{1\sim 2}$, respectively. Consider how the bootstrap is applied to obtain estimates of $\text{Var}(\hat{\Theta})$.

The simplest case is when N/n_1 is an integer, say k . First, we form the pseudo-population label set

$$U_A^* = U_{A(2)}^* \cup U_{A(1\sim 2)}^*, \quad (2.32)$$

where $U_{A(2)}^*$ consists of k copies of the units in S_2 and $U_{A(1\sim 2)}^*$ consists of k copies of the units in $S_{1\sim 2}$. We then perform the following three steps:

1. Draw a SRSWOR of size n_2 from $U_{A(2)}^*$ and denote this set by S_2^* .
2. Draw a SRSWOR of size $n_{1\sim 2}$ from $U_{A(1\sim 2)}^*$ and denote this set by $S_{1\sim 2}^*$.
3. Compute $\hat{\Theta}_1^* = \hat{\Theta}_1(S_{1\sim 2}^*, S_2^*)$ which has the same functional form as $\hat{\Theta}(S_{1\sim 2}, S_2)$, but is computed for the $n_1 = n_{1\sim 2} + n_2$ units in $S_1^* = S_{1\sim 2}^* \cup S_2^*$.

Repeat steps 1 to 3 some large number, Q , times to obtain $\hat{\Theta}_1^*, \dots, \hat{\Theta}_Q^*$. Then, an estimator of $\text{Var}(\hat{\Theta})$ is

$$\text{var}_{BSS}(\hat{\Theta}) = \sum_{q=1}^Q \frac{(\hat{\Theta}_q^* - \hat{\Theta}^*)^2}{Q-1}, \quad (2.33)$$

where $\hat{\Theta}^* = \sum_{q=1}^Q \hat{\Theta}_q^* / Q$.

Using the methods of Rao and Wu (1988), it can now be shown that $\text{var}_{BSS}(\hat{\Theta})$ is a consistent estimator of $\text{Var}(\hat{\Theta})$. If $N = kn_1 + r$, where $0 < r < n_1$, the procedure is modified as follows using the Bickel and Freedman

procedure. First, form the pseudo-population U_A^* as above consisting of kn_1 units. In addition, form the pseudo population $U_B^* = U_{B(1\sim 2)}^* \cup U_{B(2)}^*$ of size $(k+1)n_1$ where $U_{B(1\sim 2)}^*$ and $U_{B(2)}^*$ consist of $k+1$ copies of the labels in $S_{1\sim 2}$ and S_2 , respectively. Then, for αQ of the bootstrap samples, select $S_1^* = S_{1\sim 2}^* \cup S_2^*$ from U_A^* and for $(1-\alpha)Q$ samples, select S_1^* from the pseudo-population, U_B^* using the three-step procedure described above, where

$$\alpha = \left(1 - \frac{r}{n_1}\right) \left(1 - \frac{r}{N-1}\right). \quad (2.34)$$

2.4.2 Estimation of Bias and MSE

The bootstrap procedure can also provide an estimate of estimator bias. The usual bootstrap bias estimator (see Efron and Gong 1983; Rao and Wu 1988) is $b(\hat{\Theta}) = \hat{\Theta}^* - \hat{\Theta}$ where $\hat{\Theta}^* = \sum_{q=1}^Q \hat{\Theta}_q^* / Q$ and $\hat{\Theta}$ is the estimate computed from the full sample. Note that $\hat{\Theta}_q^*(q = 1, \dots, Q)$ and $\hat{\Theta}$ have the same functional form and are based upon the same model assumptions. Thus $b(\hat{\Theta})$ does not reflect the contribution to bias due to model failure. We propose an alternative estimator of bias which we conjecture is an improvement over $b(\hat{\Theta})$.

Recall from (2.4) that $\bar{B} = E(y_i - \mu_i)$ where $E(\cdot)$ denotes expectation over both the measurement error and sampling error distributions. Thus, \bar{B} may be rewritten as $\bar{B} = \sum_{i=1}^N (Y_i - \mu_i) / N$ where $Y_i = E(y_i | i)$. Since Y_i is unknown and unobservable for all $i \in U$, \bar{B} is also unknown and unobservable. Therefore, we shall construct a pseudo population resembling U , denoted by U^* , such that $\bar{B}^* = E^*(y_i - \mu_i)$ is known, where $E^*(\cdot)$ is expected value with respect to both the measurement error and the sampling distributions associated with U^* .

Let $U^* = \bigcup_{h=1}^L U_h^*$ where U_h^* consists of $k_h = N_h/n_{1h}$ copies of the units in S_{1h} . Here we have assumed k_h is an integer, but we will subsequently relax the assumption. Further, denote by y_i^* the value of the characteristic for the unit $i \in U^*$. This value is equal to the y_i for the corresponding unit in S_1 . Thus, the population total of the y_i^* is $Y^* = \sum_{i \in U^*} y_i^* = \hat{Y}_{1st}$ for \hat{Y}_{1st} defined in (2.13). Analogously, define the true value for unit $i \in U^*$ as $\mu_i^* = \mu_j$ for $i \in U^*$ corresponding to $j \in S_2$. For $j \in S_{1\sim 2}$, μ_j is unknown; however, for our pseudo-population we could generate pseudo-values for the μ_i^* such that $M^* = \sum_{i \in U^*} \mu_i^* = \hat{M}_{2st}$ where \hat{M}_{2st} is defined in (2.11). Thus, for U^* , $B^* = \hat{Y}_{1st} - \hat{M}_{2st} = \hat{B}_{12st}$ defined in (2.13). As we shall see, it is not necessary to generate the pseudo-values for μ_i^* in order to evaluate the bias in the estimators of B^* .

Note that under stratified sampling, $U^* = U_A^*$, as defined in Section 2.4. Further, the bootstrap procedure described in this section is equivalent to repeated sampling from U^* and the alternative estimators $\hat{\Theta}_1, \dots, \hat{\Theta}_p$ of B

may also be considered estimators of B^* . Since B^* is known, the bias of $\hat{\Theta}$ as an estimator of B^* is $\hat{B}^* = \hat{\Theta} - B^*$ and the corresponding MSE may be estimated as

$$\widehat{MSE} = \sum_q (\hat{\Theta}_q - B^*)^2 / Q$$

$$\doteq \text{var}_{BSS}(\hat{\Theta}) + (\hat{\Theta}^* - B^*)^2, \quad (2.35)$$

where $\text{var}_{BSS}(\hat{\Theta})$, $\hat{\Theta}_q$, and $\hat{\Theta}^*$ are defined in Section 2.4. It can be easily verified that these results still hold when k_h is non-integer.

Thus, the bootstrap procedure provides a method for evaluating the MSE of alternative estimators for estimating B^* . Further, the pseudo-population U^* is a reconstruction of U based upon copies of the values for the units in S_1 and S_2 . Thus, it is reasonable to use \hat{B}^* and \widehat{MSE}^* to evaluate alternative estimators of B .

3. APPLICATION TO THE AGRICULTURAL SURVEY

3.1 Description of the Survey

The National Agricultural Statistics Service (NASS) annually conducts a series of surveys which are collectively referred to as the Agricultural Survey (AS) program. The purpose of these surveys is to collect data related to specific agricultural commodities at the state and national levels. Each December in the years 1988-1990, reinterview studies designed to assess the measurement bias in the data collected by Computer Assisted Telephone Interviewing (CATI) were conducted in six states: Indiana, Iowa, Minnesota, Nebraska, Ohio, and Pennsylvania. The reinterview techniques employed in these three studies are very similar to those used by the U.S. Census Bureau (see, for example, Forsman and Schreiner 1991). However, unlike the Census Bureau's program, the major objective in the NASS studies is the estimation of measurement bias rather than interviewer performance evaluation.

As noted above, only AS responding units whose original interview was conducted by CATI were eligible for selection into the reinterview sample. The reasons for this restriction on sampling were primarily cost, timing, and convenience. However, a large proportion of the AS is conducted by CATI and, thus, information regarding AS measurement bias for this group would provide important information for the entire AS program.

For the NASS reinterview studies, the interviewing staff consisted of a mix of field supervisors and experienced field interviewers. This interviewing staff, which was a separate corps of interviewers from those used for CATI, conducted face-to-face reinterviews in a subsample of AS

units for a subset of AS survey items. To minimize any problems that respondents may have with recall, the reinterviews were conducted within 10 days of the original interview. Differences between the original AS and reinterview responses were reconciled to determine the "true" value. Considerable effort was expended in procedural development, training, and supervision of the reinterview process to ensure that the final reconciled response was completely accurate. For the most part, the wording of the subset of AS questions asked in the reinterview was identical to that of the parent survey. The reinterviewers attempted to contact the most knowledgeable respondent in order to ensure the accuracy of the reconciled values.

In this report, only the 1990 data are analyzed. Table 1 presents the reinterview sample sizes for this study.

Table 1
Sample Sizes by Survey Item

Item	x	y	μ
	U	S_1	S_2
All wheat stocks	108,267	8,176	1,157
Corn planted acres	225,269	8,211	1,157
Corn stocks	225,269	7,990	1,115
Cropland acreage	278,045	8,274	1,141
Grain storage capacity	207,460	8,126	1,104
Soybean planted acreage	171,761	8,211	1,156
Soybean stocks	171,761	8,113	1,130
Total land in farm	276,450	8,309	1,159
Total hog/pig inventory	248,571	8,247	1,142
Winter wheat seedings	108,267	8,211	1,150

3.2 Comparison of the Estimators of M and B

Using the December 1990 Agricultural Survey and its corresponding reinterview survey data, the estimators developed in the previous section were compared. Estimates of standard errors and mean squared errors were computed using the Bickel-Freedman bootstrap procedure described in Section 2.4, with $Q = 300$ bootstrap samples. Table 2 displays the results for six of the estimators: \hat{B}_{2st} , the traditional difference estimator; \hat{B}_{x2stR} , the weighted ratio estimator; \hat{B}_{x2stR} , the robust (outlier deletion) version of \hat{B}_{x2stR} ; \hat{B}_{SSW} , the Särndal, Swensson and Wretman estimator; \hat{B}_M , the unweighted model-based estimator; and \hat{B}_M , the robust (outlier deletion) version of \hat{B}_M .

3.3 Summary of Results

Table 2 presents a summary of the results from our study. The first data column is the known value of $B^* = E(y_i^* - \mu_i^*)$, the bias parameter for the pseudo-population, U^* . The other data columns contain the values of the estimators with their standard errors in parentheses, where $\text{s.e.}(\hat{\theta}) = \sqrt{\text{var}_{BSS}(\hat{\theta})}$. The last four rows of the table correspond, respectively, to:

- the number of items (out of 10) for which a 95% confidence interval contains B^* ;
- the average coefficient of variation (C.V.);
- the average square root of $\widehat{\text{MSE}}$ (RMSE); and
- the average absolute relative bias.

A striking feature of these results is the large disparity among the six estimators across all commodities; particularly for All Wheat Stocks. For this commodity, the range of estimates is -94.2 to 103.2 . Also indicated (by

the ‡ symbol) in Table 2 is whether a 95% confidence interval, *i.e.*, $[\hat{\theta} - 2\text{s.e.}(\hat{\theta}), \hat{\theta} + 2\text{s.e.}(\hat{\theta})]$, covers the parameter B^* . The best performer for parameter coverage is \hat{B}_{SSW} which produced confidence intervals that covered B^* for eight out of ten commodities. \hat{B}_{2st} was the next best with six and \hat{B}_M was third with five. The traditional ratio estimator and its robust version were the worst performers with only one commodity having a confidence interval covering B^* .

Application of the mean squared error criterion presents a different picture. Here, \hat{B}_M emerged as the estimator having the smallest average root MSE. However, \hat{B}_{SSW} and \hat{B}_{2st} are not much greater. Further, \hat{B}_{SSW} was the estimator having the smallest average absolute relative bias. Only two commodities were estimated with significant biases using this estimator. Thus, it appears from these results that \hat{B}_{SSW} is the preferred estimator using overall performance as the evaluation criterion.

Table 2
Comparison of Estimators with, B^* , the Pseudo-Population Value of the Bias†

Characteristic	B^*	\hat{B}_{2st}	\hat{B}_{x2stR}	\hat{B}_{x2stR}	\hat{B}_{SSW}	\hat{B}_M	\hat{B}_M
All wheat stocks	42.3	-6.1 (12.3)	103.2 (17.6)	-94.2 (16.5)	-0.9‡ (24.8)	19.2‡ (16.5)	10.6‡ (16.7)
Corn planted acreage	-1.8	1.1‡ (1.1)	11.7 (1.3)	10.1 (1.1)	0.3‡ (1.2)	-4.7‡ (1.9)	-5.0 (1.5)
Corn stocks	-6.4	-5.4‡ (1.5)	2.4 (1.6)	0.2 (1.3)	-6.5‡ (1.6)	-7.9‡ (2.4)	-9.3‡ (2.2)
Cropland acreage	27.0	-19.6 (8.3)	-15.0 (8.3)	7.0 (3.1)	-19.6 (8.2)	-36.8 (11.0)	-12.8 (4.0)
Grain storage capacity	-3.37	1.4‡ (3.7)	32.3 (3.7)	29.5 (2.6)	-0.1‡ (3.9)	-6.9 (3.0)	-6.8 (2.5)
Soybean planted acreage	-4.4	0.8 (0.8)	13.0 (1.0)	9.9 (0.9)	-0.3 (1.0)	-2.9 (1.1)	-2.7 (1.0)
Soybean stocks	-0.01	2.8‡ (3.1)	21.3 (2.9)	5.0 (2.3)	0.2‡ (3.5)	-11.0 (3.6)	-8.8 (3.4)
Total land in farm	-20.0	-24.7‡ (10.4)	-18.8‡ (12.5)	-2.6 (7.6)	-25.7‡ (10.7)	-44.5‡ (13.4)	-21.2 (5.8)
Total hogs/pigs inventory	-0.1	-2.1 (0.9)	3.4 (1.1)	-0.0‡ (1.0)	-2.2‡ (1.1)	-2.5‡ (1.3)	-1.6‡ (1.0)
Winter wheat seedings	-0.6	-0.5‡ (0.4)	3.8 (0.6)	1.8 (0.5)	-1.2‡ (0.6)	1.1 (0.4)	1.1 (0.4)
Number of items where C.I. covers B^*		6	1	1	8	5	3
Average C.V.		1.01	.30	11.1	9.5	.41	.48
Average RMSE		13.2	22.4	25.2	12.9	14.9	10.8
Average Relbias		30.8	220.0	53.4	4.9	113.1	91.3

† Standard errors in parentheses.

‡ 95% confidence interval covers the pseudo population parameter.

4. CONCLUSIONS AND RECOMMENDATIONS

In this article, we developed a general methodology for constructing and evaluating weighted and unweighted model prediction estimators of measurement bias for stratified random, two-phase sample designs. The proposed estimators incorporate information on the observations, y , from the first phase sample, and an auxiliary variable, x . Model robust versions of the estimators were also considered and evaluated. The ultimate goal of model prediction estimation is to identify estimators which make "optimal" use of the data (y, μ, x) . The general estimation and evaluation methodology for achieving this goal was illustrated for the ordinary regression model with no intercept. However, the methodology can be easily extended to multivariate, intercept models.

Our proposed evaluation criteria are based upon estimates of bias, variance, and mean squared error computed using a bootstrap resampling methodology. The method of Bickel and Freedman was extended to two-phase sampling for this purpose. It was shown both analytically and empirically that the usual NDR estimator is not optimal under the model prediction approach to estimating measurement bias. Our analyses found that, for the six estimators we considered, the estimator derived from the work of Särndal *et al.* (1992), was the best overall estimator by the bootstrap evaluation criteria.

Incorporating auxiliary information into the estimation of measurement bias creates a number of practical problems which may increase the costs and reduce the timeliness of producing the estimates. First, the auxiliary variable, x , must be available, at least in aggregate form, for all socioeconomic and geographic domains for which model prediction estimates are desired. This could be a large data management task. Further, the complexity of the variance estimator using analytical methods increases with the complexity of the bias estimator. Although simpler, the bootstrap variance estimation method can be prohibitively expensive if computer time must be purchased. However, these difficulties are not insurmountable, especially if a high-powered microcomputer is available. Further, given the cost of reinterview surveys for estimating measurement bias, even moderate increases in precision in the bias estimators can result in substantial cost savings.

The model prediction approach has the potential for extracting the maximum information on response bias from reinterview surveys and thus model prediction estimators will usually be more efficient than the traditional net difference estimator. In addition, the model prediction approach may also offer a means for extrapolating estimates of bias to areas which were not sampled. As an example, in the NASS application, the reinterview sample was drawn only from the CATI areas for reasons of operational convenience and cost efficiency. However, by using prediction models which are functions of the

original responses and other available characteristics, it may be possible to predict the measurement bias in the non-CATI survey areas from the local characteristics of these areas – a type of "synthetic" estimation. Although this application of model-based estimation was not considered in this paper, it is a natural extension of the methodology and one which will be evaluated in a subsequent study.

Also for future research, we intend to incorporate multivariate, intercept models in the estimation of measurement bias. Since the bootstrap evaluation criteria developed in this article are general, no changes in the evaluation methodology are required to handle the addition of variables in the estimation models. Further, the model assumptions and the methods for handling outliers will be refined and evaluated in a subsequent paper. Finally, we need to explore the effect on estimation of departures from the model assumptions, particularly the assumption that the reinterview observation is without error. As Fuller (1991) has shown, if the reinterview is fallible but unbiased, the variance of the predicted values increases but the predictions are still unbiased. Thus, under these assumptions, one could explore the relative precision of the alternative estimators of measurement bias in order to determine the robustness of the model prediction approach.

ACKNOWLEDGEMENTS

The authors thank Manuel Cárdenas for his valuable assistance and the referees for their helpful comments.

REFERENCES

- BICKEL, P., and FREEDMAN, D.A. (1984). Asymptotic normality and the bootstrap in stratified sampling. *The Annals of Statistics*, 12, 470-482.
- BIEMER, P., and STOKES, L. (1991). Approaches to the modeling of measurement errors. In *Measurement Errors in Surveys*, (Eds. P. Biemer, *et al.*). New York: John Wiley and Sons.
- COCHRAN, W. (1977). *Sampling Techniques*. New York: John Wiley and Sons.
- EFRON, B., and GONG, G. (1983). A leisurely look at the bootstrap, the jackknife, and cross-validation. *The American Statistician*, 31, 36-48.
- FORSMAN, G., and SCHREINER, I. (1991). The design and analysis of reinterview: an overview. In *Measurement Errors in Surveys*, (Eds. P. Biemer, *et al.*). New York: John Wiley and Sons.
- FULLER, W.A. (1991). Regression estimation in the presence of measurement error. In *Measurement Errors in Surveys*, (Eds. P.P. Biemer, *et al.*). New York: John Wiley and Sons, 617-636.

- HANSEN, M., MADOW W., and TEPPING, B. (1983). An evaluation of model-dependent and probability sampling inferences in sample surveys. *Journal of the American Statistical Association*, 78, 776-793.
- RAO, J.N.K., and WU, C. (1988). Resampling inference with complex survey data. *Journal of the American Statistical Association*, 83, 231-241.
- ROYALL, R., and HERSON, J. (1973). Robust estimation in finite populations I. *Journal of the American Statistical Association*, 68, 880-893.
- ROYALL, R., and CUMBERLAND, W. (1978). Variance estimation in finite population sampling. *Journal of the American Statistical Association*, 73, 351-361.
- ROYALL, R., and CUMBERLAND, W. (1981). The finite-population linear regression estimator and estimators of its variance – an empirical study. *Journal of the American Statistical Association*, 76, 924-930.
- SÄRNDAL, C.-E., SWENSSON, B., and WRETMAN, J. (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.

Model-Based Estimation of Record Linkage Error Rates

J.B. ARMSTRONG and J.E. MAYDA¹

ABSTRACT

Record linkage is the matching of records containing data on individuals, businesses or dwellings when a unique identifier is not available. Methods used in practice involve classification of record pairs as links and non-links using an automated procedure based on the theoretical framework introduced by Fellegi and Sunter (1969). The estimation of classification error rates is an important issue. Fellegi and Sunter provide a method for calculation of classification error rate estimates as a direct by-product of linkage. These model-based estimates are easier to produce than the estimates based on manual matching of samples that are typically used in practice. Properties of model-based classification error rate estimates obtained using three estimators of model parameters are compared.

KEY WORDS: Mixture model; Latent variable model; Iterative scaling.

1. INTRODUCTION

Computer files containing information about individuals, businesses or dwellings are used in many statistical applications. The linking of records that refer to the same entity is often required. The process of linking records referring to the same entity is called exact matching. If all records involved in an application have been accurately assigned a unique identifier, exact matching is trivial. Record linkage methods deal with the problem of exact matching when a unique identifier is not available. In that case, each record typically includes a number of data fields containing identifying information that could be used for matching. Problems in matching are due to errors in these data or due to the same value for a particular field being valid for more than one entity.

Applications of record linkage include the unduplication of lists of dwellings or businesses obtained from various sources to create survey frames. In addition, record linkage is widely used in applications related to health and epidemiology. Work in this area typically involves matching records containing information on individuals in industrial or occupational cohorts to records documenting the illness or death of individuals. For example, record linkage methodology for follow-up studies of persons exposed to radiation is discussed in Fair, Newcombe and Lalonde (1988).

The record linkage problem can be formulated using two data files that correspond to two populations. Each file may contain information for all entities in the corresponding population or information for a random sample of entities. The file A contains N_A records and the file B contains N_B records. The set of record pairs formed as the cross-product of A and B is denoted by $C = \{(a, b);$

$a \in A, b \in B\}$. C contains $N = N_A \cdot N_B$ record pairs. The objective of record linkage is to partition the set C into two disjoint sets – the set of true matches, denoted by M , and the set of true non-matches, U .

The theoretical framework introduced by Fellegi and Sunter (1969) is the basis of a great deal of applied work. For each record pair, a decision is taken concerning whether or not the records refer to the same entity after examining data recorded on files A and B . The possible decisions are link (A_1), non-link (A_3) and possible link (A_2). There are two types of errors. First, decision A_1 may be taken for a record pair that is a member of U , the set of true non-matches. Second, decision A_3 may be taken for a record pair that is a member of set M , the set of true matches. Acceptable levels of classification error are specified before the files are linked. A record pair is classified as a possible link if the data do not provide sufficient evidence to justify classification of the pair as a link or non-link at error levels less than or equal to those specified. Accurate estimation of classification error rates associated with various decision rules is necessary to determine an appropriate rule. The classification error rate for true non-matches is $P(A_1 | U)$. The error rate for true matches is $P(A_3 | M)$.

Estimates of classification error rates can be obtained by selecting a sample of record pairs from the set C and manually determining the true match status of sampled pairs. Applications of this approach are described in Bartlett *et al.* (1993). Sampling may be both costly and cumbersome to implement, particularly when the same linkage must be done for a number of pairs of files, each with slightly different characteristics. Belin and Rubin (1991) describe another method of error rate estimation

¹ J.B. Armstrong and J.E. Mayda, Statistics Canada, Business Survey Methods Division, 11–RH Coats Bldg, Tunney's Pasture, Ottawa, Ontario, K1A 0T6.

that requires true match status for record pairs in a pilot study. In contrast to the straightforward sampling approach, the Belin-Rubin method provides a framework for the application of information obtained from the pilot study to larger linkages involving similar data.

The Fellegi-Sunter framework provides a method for calculation of error rate estimates using estimates of probabilities that record pairs will agree on various combinations of data fields. Calculation of these model-based error rate estimates is straightforward and manual determination of the true match status of record pairs is not required. However, they often have poor properties in applied work. See, for example, Belin (1990). In this paper, the potential for improvement of the properties of model-based error rate estimates through careful estimation of agreement probabilities is examined.

Three alternative estimation methods are evaluated. The approaches described use only the information on files A and B. They do not rely on auxiliary information. Model-based error rate estimates obtained using each alternative are compared with actual error rates using both synthetic data that incorporate important characteristics of data from health applications of record linkage, and information from an actual record linkage application.

The plan of the paper is as follows. Section 2 includes details of the model-based classification error rate estimation method introduced by Fellegi and Sunter. The model for agreement probabilities that forms the basis of subsequent discussion of estimation methods is also specified. Two estimation methods that rely on an important independence assumption are described in Section 3. A third alternative that does not require independence is discussed in Section 4. The results of comparisons of the three approaches using synthetic data are reported in Section 5. The results of evaluation work with information from a real application are described in Section 6. Section 7 contains some concluding remarks.

2. THEORETICAL CONCEPTS

Relevant aspects of the theory for record linkage developed by Fellegi and Sunter (1969) are summarized in this section. In the Fellegi-Sunter framework, estimates of classification error rates are calculated using estimates of probabilities of agreement on various combinations of data fields. Applications of the theory of Fellegi and Sunter usually involve the assumption that the probability that a record pair will agree on a particular data field is independent of the results of comparisons for other fields. The theory is nevertheless very flexible, allowing for any pattern of dependence between results of comparisons for different data fields. A parameterization of dependence in terms of loglinear effects is given.

2.1 Model-Based Classification Error Rate Estimation

To obtain information related to the classification of a record pair as a link (A_1), non-link (A_3) or possible link (A_2), data fields containing identifying information are compared. In an application involving records referring to persons, separate comparisons of family names, given names, and dates of birth might be performed. The outcome of a comparison is a numerical code representing a statement like "names agree", "names disagree", "name missing on one or both files", "names agree and both are George" or "names disagree but their first two characters agree". The outcome codes used in applied work differ between applications and between comparisons in the same application. The smallest number of outcome codes that can be used for any comparison is two – corresponding to agreement and disagreement. An outcome code corresponding to "missing on one or both files" is usually needed in applied work. The agreement outcome may be replaced by a number of value-specific outcomes (such as "names agree and both are George"). Certain disagreements may be coded as partial agreements (such as "names disagree but their first two characters agree").

For present purposes, we consider agreement and disagreement outcomes only. In the case of K matching fields, we introduce the outcome vector $\underline{x}^j = (x_1^j, x_2^j, \dots, x_K^j)$ for record pair j . We have $x_k^j = 1$ if record pair j agrees on data field k and $x_k^j = 0$ if record pair j disagrees on data field k .

Newcombe *et al.* (1959) introduced the idea that decisions concerning whether or not a pair of records represent the same entity should be based on the ratio

$$R(\underline{x}) = P(\underline{x} | M) / P(\underline{x} | U), \quad (1)$$

where $\underline{x} = (x_1, x_2, \dots, x_K)$ is the generic outcome vector, $P(\underline{x} | M)$ is the probability that comparisons for a record pair that is a true match will produce outcome vector \underline{x} , and $P(\underline{x} | U)$ is the probability of \underline{x} for a record pair that is a true non-match. The optimality of record linkage methods involving this ratio was demonstrated by Fellegi and Sunter.

In the Fellegi-Sunter framework, a linkage rule assigns a probability of each classification decision (A_1 , A_2 and A_3) to each outcome vector. The decision function corresponding to outcome vector \underline{x} is $d(\underline{x}) = (P(A_1 | \underline{x}), P(A_2 | \underline{x}), P(A_3 | \underline{x}))$. Acceptable rates of classification error for true non-matches and true matches are specified before linkage is conducted. We denote these pre-specified error rates by μ and λ respectively. Among the class of record linkage rules satisfying the relations $P(A_1 | U) \leq \mu$ and $P(A_3 | M) \leq \lambda$ for fixed values of μ and λ , Fellegi and Sunter define the optimal linkage rule as the rule that minimizes $P(A_2)$, the probability that a record pair will be classified as a possible link. The optimal rule has the form

$$\begin{aligned}
d(\underline{x}^j) &= (1,0,0) \quad \text{if } \omega^j > \tau_1 \\
d(\underline{x}^j) &= (P_\mu, 1 - P_\mu, 0) \quad \text{if } \omega^j = \tau_1 \\
d(\underline{x}^j) &= (0,1,0) \quad \text{if } \tau_2 < \omega^j < \tau_1 \\
d(\underline{x}^j) &= (0,1 - P_\lambda, P_\lambda) \quad \text{if } \omega^j = \tau_2 \\
d(\underline{x}^j) &= (0,0,1) \quad \text{if } \omega^j < \tau_2
\end{aligned} \tag{2}$$

where $\tau_1 \geq \tau_2$, the “weight” ω^j is defined as $\omega^j = \log(R(\underline{x}^j))$ and P_μ and P_λ are positive constants in the interval $[0,1)$. (Refer to Fellegi and Sunter (1969) for full details.) Determination of τ_1 and τ_2 requires the estimation of classification error rates corresponding to various choices for these threshold values, underscoring the importance of accurate estimation of classification error rates in the Fellegi-Sunter framework.

Model-based estimates of classification error rates can be calculated using estimates of outcome probabilities for true matches and true non-matches. Let $\hat{P}(\underline{x} | M)$ and $\hat{P}(\underline{x} | U)$ denote estimates of the probabilities of outcome vector \underline{x} for true matches and true non-matches and denote the ratio of these estimates by $\hat{R}(\underline{x})$. The model-based estimate of the classification error rate for true matches based on decision rule (2) is

$$\hat{\lambda} = \sum_{\underline{x} \in L(\tau_2)} \hat{P}(\underline{x} | M) + P_\lambda \sum_{\underline{x} \in Q(\tau_2)} \hat{P}(\underline{x} | M) \tag{3}$$

where $L(\tau_2) = \{\underline{x}; \log(\hat{R}(\underline{x})) < \tau_2\}$ and $Q(\tau_2) = \{\underline{x}; \log(\hat{R}(\underline{x})) = \tau_2\}$.

The model-based estimate of the classification error rate for true non-matches is

$$\hat{\mu} = \sum_{\underline{x} \in G(\tau_1)} \hat{P}(\underline{x} | U) + P_\mu \sum_{\underline{x} \in Q(\tau_1)} \hat{P}(\underline{x} | U) \tag{4}$$

where $G(\tau_1) = \{\underline{x}; \log(\hat{R}(\underline{x})) > \tau_1\}$ and $Q(\tau_1) = \{\underline{x}; \log(\hat{R}(\underline{x})) = \tau_1\}$.

2.2 A Model For Outcome Probabilities

Calculation of model-based classification error rate estimates requires estimation of $P(\underline{x} | M)$ and $P(\underline{x} | U)$ for each of the 2^K possible values of \underline{x} . The probability density function for \underline{x} is a mixture of two probability densities given by

$$f(\underline{x}) = pP(\underline{x} | M) + (1 - p)P(\underline{x} | U), \tag{5}$$

where p is the probability that a record pair chosen at random is a true match. The outcome probabilities depend on the frequency distributions of identifiers for entities represented on files A and B, as well as the probabilities

that errors are introduced when identifiers are recorded on the files. Fellegi and Sunter (1969, pp. 1192-1194) describe a method of estimating agreement probabilities involving their definition in terms of frequency distributions and error probabilities. They recommend use of the method when prior information is available.

In the present paper, we consider situations in which the data on files A and B and the outcome vectors \underline{x}^j , $j = 1, 2, \dots, N$, represent the only information available for estimation of outcome probabilities. A loglinear structure for the outcome probabilities is the most general parameterization. The saturated loglinear model for outcome probabilities for true matches is

$$\begin{aligned}
\log(P(\underline{x} | M)) &= M(0) + M(1)_{x_1} + M(2)_{x_2} + \dots \\
&+ M(K)_{x_K} + M(1)M(2)_{x_1, x_2} + \dots \\
&+ M(K-1)M(K)_{x_{K-1}, x_K} + \dots \\
&+ M(1)M(2)\dots M(K)_{x_1, x_2, \dots, x_K}, \tag{6}
\end{aligned}$$

with the usual restrictions

$$\sum_{x_J} M(J)_{x_J} = 0, \quad J = 1, 2, \dots, K,$$

$$\begin{aligned}
\sum_{x_{J_1}} M(J_1)M(J_2)_{x_{J_1}, x_{J_2}} &= \sum_{x_{J_2}} M(J_1)M(J_2)_{x_{J_1}, x_{J_2}} = 0, \\
&\forall J_1, J_2, \quad \text{etc.},
\end{aligned}$$

as well as the restriction

$$\sum_x P(\underline{x} | M) = 1.$$

The saturated model for $P(\underline{x} | U)$ is analogous.

If saturated loglinear models for $P(\underline{x} | M)$ and $P(\underline{x} | U)$ are employed, the density function includes $2^{K+1} - 1$ unknown parameters. It is not possible to identify all these parameters when no auxiliary information is available. In order to obtain a model that can be identified and to simplify the estimation problem, the assumption that the outcomes of comparisons for different data fields are independent is often employed. Under the assumption of independence, we denote the probabilities of agreement among record pairs that are true matches and true non-matches, respectively, by

$$m_k = P(x_k = 1 | M), \quad k = 1, 2, \dots, K,$$

$$u_k = P(x_k = 1 | U), \quad k = 1, 2, \dots, K.$$

Outcome probabilities can be written as

$$P(\underline{x} | M) = \prod_{k=1}^K m_k^{x_k} (1 - m_k)^{(1-x_k)},$$

$$P(\underline{x} | U) = \prod_{k=1}^K u_k^{x_k} (1 - u_k)^{(1-x_k)}.$$

This model involves $2 \cdot K + 1$ unknown parameters, namely $(\underline{m}, \underline{u}, p)$, where $\underline{m} = (m_1, m_2, \dots, m_K)$, $\underline{u} = (u_1, u_2, \dots, u_K)$. There are, of course, a number of intermediate models between the saturated model and the independence model. Methods that can be used to estimate the independence model are described in Section 3. Estimation of intermediate models is discussed in Section 4.

3. ESTIMATION UNDER INDEPENDENCE ASSUMPTION

3.1 Method of Moments

A methods of moments estimator of $P(\underline{x} | M)$ and $P(\underline{x} | U)$ can be employed in the case of independence. The estimator is based on a system of $2 \cdot K + 1$ equations that provide expressions for functionally independent moments of \underline{x} in terms of the parameters. The equations are

$$E\left(\prod_{k \neq i}^K x_k\right) = pN \prod_{k \neq i}^K m_k + (1 - p) N \prod_{k \neq i}^K u_k,$$

$$i = 1, 2, \dots, K$$

$$E(x_i) = pNm_i + (1 - p) Nu_i, \quad i = 1, 2, \dots, K,$$
(7)

$$E\left(\prod_{k=1}^K x_k\right) = pN \prod_{k=1}^K m_k + (1 - p) N \prod_{k=1}^K u_k.$$

To obtain estimates of the parameters using the method of moments, it is necessary to solve the equations after expectations have been replaced by averages calculated using record pairs in C . The equation system for $K = 3$ was given by Fellegi and Sunter, who also derived a closed form solution that exists if some mild conditions are satisfied. Their paper included a word of caution concerning use of the method in the case of departures from independence. For $K > 3$, a closed form solution is not available but standard numerical methods can be used. Parameter estimates obtained using the method of moments are statistically consistent if the independence assumption is true.

3.2 Iterative Method

The iterative method was developed by record linkage practitioners. Although the method is not based on the probability distribution of the outcome vector, it does

make use of the independence assumption. Application of the iterative method is described by several authors, including Newcombe (1988). Statistics Canada's record linkage software, CANLINK, is set up to facilitate use of the iterative method.

The method requires initial estimates of the agreement probabilities for true matches and non-matches. For true matches, guesses based on previous experience must be employed. To obtain initial estimates of agreement probabilities among record pairs that are true non-matches it is typically assumed that these probabilities are equal to the probabilities of agreement among record pairs chosen at random, namely that,

$$u_k = P(x_k = 1), \quad k = 1, 2, \dots, K.$$

Suppose that $J(k)$ different values for data field k appear on file A and/or file B. Denote the frequencies of these values on file A by $f_{k1}, f_{k2}, \dots, f_{kJ(k)}$ and denote the file B frequencies by $g_{k1}, g_{k2}, \dots, g_{kJ(k)}$. For a particular value one, but not both, of the counts may be zero. The initial estimate of u_k is

$$\hat{u}_k^0 = \sum_{j=1}^{J(k)} (f_{kj} g_{kj}) / N. \quad (8)$$

Given these probability estimates, initial sets of matches and non-matches, denoted by M^0 and U^0 respectively, are obtained using a decision rule

$$j \in M^0 \quad \text{if} \quad \omega^j > \tau_1^0,$$

$$j \in U^0 \quad \text{if} \quad \omega^j < \tau_2^0.$$

Next, frequency counts among record pairs in the sets M^0 and U^0 are used as new estimates of agreement probabilities. These estimates are used to obtain new sets of matches and non-matches and the iterative process is continued until consecutive estimates of agreement probabilities are sufficiently close.

In most applications, the assumption that the probability of agreement among record pairs that are true non-matches is equal to the probability of agreement among all record pairs is a good one and iteration does not lead to any important changes in estimates of non-match agreement probabilities. However, the first iteration often produces large changes in agreement probability estimates for true matches. Typically, there are no substantial changes at the second iteration.

It should be noted that the statistical properties of the iterative method are unclear. In practice, performance of the method will depend on the choice of the initial thresholds τ_1^0, τ_2^0 . These thresholds are typically chosen subjectively. The simulations reported in Section 5 provide information about the effects of various initial thresholds.

4. RELAXING THE INDEPENDENCE ASSUMPTION – ESTIMATION USING ITERATIVE SCALING

Methods of estimation for latent variable models can be used to estimate agreement probabilities when the dependence between outcomes of comparisons for different matching fields is parameterized in terms of loglinear effects. Winkler (1989) and Thibaudeau (1989) have estimated agreement probabilities using loglinear models including all interaction terms up to third or fourth order to parameterize dependencies. The formulation presented here facilitates use of loglinear models including selected interactions. Match status can be considered a latent variable with two levels (true match and true non-match). Let $c_{0,\underline{x}}$ and $c_{1,\underline{x}}$ denote the numbers of true non-matches and true matches, respectively, with outcome vector \underline{x} in a record linkage application involving K matching variables. These counts are, of course, unobservable since the value of the latent variable for each record pair is unknown. Instead, $c_{\underline{x}} = c_{0,\underline{x}} + c_{1,\underline{x}}$ is observed.

Using the parameterization of dependence in terms of loglinear effects and a saturated model for true matches, we can write

$$\begin{aligned} \log(c_{1,\underline{x}}/Np) = & M(0) + M(1)_{x_1} + M(2)_{x_2} + \dots \\ & + M(K)_{x_K} + M(1)M(2)_{x_1,x_2} + \dots \\ & + M(K-1)M(K)_{x_{K-1},x_K} + \dots \\ & + M(1)M(2) \dots M(K)_{x_1,x_2, \dots, x_K}, \end{aligned}$$

with the usual restrictions. A similar expression for true non-matches is available. The latent variable model corresponding to these saturated loglinear models is

$$\begin{aligned} \log(c_{s,\underline{x}}/w_s) = & G(0) + Z_s + G(1)_{x_1} + \dots \\ & + G(K)_{x_K} + ZG(1)_{s,x_1} + \dots + ZG(K)_{s,x_K} \\ & + \dots + G(1)G(2) \dots G(K)_{x_1,x_2, \dots, x_K} \\ & + ZG(1)G(2) \dots G(K)_{s,x_1,x_2, \dots, x_K}, \end{aligned}$$

where the index s has value zero for true non-matches and one for true matches, $w_0 = (1 - p)N$ and $w_1 = pN$. The parameters are analogous to the parameters of a saturated loglinear model for a contingency table of dimension 2^{K+1} . The usual restrictions apply. For example, the term $ZG(1)_{s,x_1}$ represents the interaction of the latent variable and the first matching variable and

$$\sum_s ZG(1)_{s,x_1} = \sum_{x_1} ZG(1)_{s,x_1} = 0.$$

This model conforms to the general latent variable model of Haberman (1979, p. 561). Additional restrictions must be imposed to identify and estimate the parameters. For simplicity, we will consider only hierarchical models. In addition, we restrict attention to models that allow all non-zero effects to interact with the latent variable.

In subsequent discussion we will denote latent variable models using symbols $G(1), G(2), \dots$, loglinear models for true matches using $M(1), M(2), \dots$ and loglinear models for true non-matches using $U(1), U(2), \dots$. In the case of four matching variables, for example, the model $G(1)G(2), G(3), G(4)$ is a latent variable model including a general level term, main effects for all four matching variables and a term for the interaction of matching variables one and two, as well as a main effects term for the latent variable (the interaction of the general level term and the latent variable), terms for the interaction of each matching variable and the latent variable and a term for the interaction of matching variables one and two and the latent variable. The model includes 12 parameters that must be estimated. The number of parameters that must be estimated in one of the latent variable models considered here is twice the number of parameters in the corresponding loglinear model.

The iterative scaling method of Haberman (1976) can be used to estimate latent variable models. The Haberman estimation method operates by raking tables that contain estimated counts for each outcome among true matches and true non-matches. Denote the estimated counts for outcome vector \underline{x} after i iterations of the Haberman algorithm by $\hat{C}_{1,\underline{x}}^i$ and $\hat{C}_{0,\underline{x}}^i$ for true matches and true non-matches, respectively. Starting values $\hat{C}_{1,\underline{x}}^0$ and $\hat{C}_{0,\underline{x}}^0$ can be constructed using estimates of agreement probabilities and the proportion of true matches obtained under the independence assumption. Each iteration of the algorithm involves a series of raking operations on the current table for true matches and the analogous rakes on the current table for true non-matches. Using the notation for hierarchical models introduced above, a set of raking operations is performed for each of the interaction terms that define the model. For four matching variables and the model $G(1)G(2), G(3)G(4)$, two sets of raking operations are performed – one for the $G(1)G(2)$ interaction and a second for the $G(3)G(4)$ interaction. For each iteration, one raking operation is performed for every level of the corresponding classification variable. Let S_{gl} denote the set of outcome vectors at level l of term g . The raking operation on the table of true matches at iteration i for level l of term g involves computation of

$$\gamma_{1,\underline{x}} = c_{\underline{x}} \hat{c}_{1,\underline{x}}^{i-1} / (\hat{c}_{1,\underline{x}}^{i-1} + \hat{c}_{0,\underline{x}}^{i-1}),$$

$$\hat{c}_{1,\underline{x}}^i = \hat{c}_{1,\underline{x}}^{i-1} \sum_{\underline{x} \in S_{gl}} \gamma_{1,\underline{x}} / \sum_{\underline{x} \in S_{gl}} \hat{c}_{1,\underline{x}}^{i-1}, \quad \forall \underline{x} \in S_{gl}.$$

The algorithm is terminated when changes between estimated counts for consecutive iterations are smaller than a given tolerance.

Haberman (1976) notes that the iterative scaling algorithm may converge to a local maximum of the likelihood function rather than to the maximum likelihood estimate. Experiments with different starting values using data sets employed in the evaluation reported in Section 5 did not yield any examples of this problem.

5. COMPARISON OF ESTIMATION METHODS - SYNTHETIC DATA

In this section, the results of comparisons of the estimation methods described in Section 3 and Section 4 are presented. The comparisons involved application of each approach to a series of synthetic data sets generated using Monte Carlo methods.

Synthetic data records containing four personal identifiers (family name, middle initial, given name, date of birth) were employed. Information on possible values of each identifier, as well as their relative frequencies, was taken from the Canadian Mortality Data Base for 1988. This database, which is frequently used in health applications of record linkage, contains a separate record for each individual death.

The independence assumption was violated among true matches in each synthetic data set. Information on the frequency of outcome vectors for true matches obtained from various record linkage projects conducted by the Canadian Center for Health Information at Statistics Canada was used during data generation. Most of the projects involved matching a cohort file to the Canadian Mortality Data Base. The frequency of each outcome vector among the true matches is shown in Table 1. The dependence in these data is clear. Although approximately 88.3% of the true matches agree on given name, the probability of agreement on given name given disagreement on middle initial and agreement on family name and birth year is only 381/1366 – about 27.9%. The value of the likelihood ratio test statistic for the independence hypothesis is 3604. This value is very extreme relative to the chi-square reference distribution with 10 degrees of freedom. (Note that one degree of freedom is lost due to the zero count for the cell (1,0,0,0).)

For each synthetic data set, file A records were generated by selecting identifiers according to relative frequencies in the 1988 Canadian Mortality Data Base. In order to simplify the data generation process, the choice of family names was restricted to the 100 most common non-francophone family names and the 100 most common francophone family names found on the 1988 file. The choice of given name was restricted to the 50 most common francophone given names and the 50 most common non-francophone

given names. All name choices excluded typographical variations. All middle initials and birth years found on the 1988 file were considered. Records with anglophone given names were more likely to receive an anglophone family name than records with francophone given names (reflecting the distribution of names in the Canadian population). Otherwise, identifiers were selected independently.

Table 1
Outcome Frequencies, Set of True
Matches, Synthetic Data

Outcome by Identifier: 0 = Disagreement, 1 = Agreement				Frequency	
Given Name	Middle Initial	Family Name	Birth Year	Count	Percent- age
0	0	0	0	7	0.03
0	0	0	1	33	0.12
0	0	1	0	125	0.45
0	0	1	1	985	3.54
0	1	0	0	5	0.02
0	1	0	1	39	0.14
0	1	1	0	202	0.73
0	1	1	1	1,848	6.65
1	0	0	0	0	0.0
1	0	0	1	13	0.05
1	0	1	0	50	0.18
1	0	1	1	381	1.37
1	1	0	0	44	0.16
1	1	0	1	451	1.62
1	1	1	0	1,751	6.30
1	1	1	1	21,860	78.65
Total				27,794	100

The starting point for file B was an exact copy of file A. Each file B record was a true match with exactly one file A record. To introduce dependence among true matches, an outcome vector was drawn from the frequency distribution in Table 1 for each file B record. Identifiers corresponding to zeroes in the outcome vector were re-selected. Consequently, the set of outcome vectors for true matches was a sample from the Table 1 distribution. The synthetic data sets also included mild departures from the independence assumption for true non-matches since the selection of given and family names was not completely independent.

Each set of simulation results reported subsequently is based on 50 Monte Carlo trials. Each trial involved generation of files A and B of size 500, estimation of \underline{m} and \underline{u} , determination of thresholds corresponding to various

model-based classification error rate estimates and calculation of actual error rates corresponding to the thresholds. The same series of 50 synthetic data sets was used for each set of simulations. Note that the set C contains 250,000 record pairs including 249,500 true non-matches for each Monte Carlo trial. In order to reduce computing time required by the simulations, only 49,500 true non-matches were used for each trial. (A small scale test was conducted to verify that reducing the number of true non-matches had a negligible effect on the estimated agreement probabilities.) True non-matches were removed from C by dividing files A and B into five corresponding blocks of size 100 and excluding record pairs involving records from blocks that did not correspond.

The method of moments equation system was solved using a variation of Newton's method that is described in detail in Moré *et al.* (1980). Computer code from IMSL (1987) was employed. Agreement probabilities of 0.9 for true matches and 0.1 for true non-matches for all matching fields were used as starting values for the solution of the equation system. The method did not appear sensitive to starting values.

The properties of the iterative method depend on the definitions of the initial sets of matches and non-matches, M^0 and U^0 . Recall that, given initial probabilities, record pairs are classified according to

$$j \in M^0 \text{ if } \omega^j > \tau_1^0,$$
$$j \in U^0 \text{ if } \omega^j < \tau_2^0.$$

When the iterative method was implemented for the simulations reported here, τ_2^0 was set equal to τ_1^0 . For each Monte Carlo trial, τ_1^0 was determined such that

$$\hat{P}(j \in U \mid \omega^j > \tau_1^0) + \gamma \cdot \hat{P}(j \in U \mid \omega^j = \tau_1^0) = \mu^0,$$

for some $\gamma \in [0,1)$, where the estimated probabilities are based on the initial iterative estimates of \underline{u} . Record pairs with weight τ_1^0 were classified in M^0 with probability γ . That is, the initial set of matches used by the iterative method was chosen to correspond to an estimated classification error rate of μ^0 for true non-matches. Starting values for m_k , $k = 1, 2, \dots, 4$, were set to 0.9.

The zero count in Table 1 (agreement on given name, disagreement on all other identifiers) was treated as a structural zero during data generation. Among loglinear models involving no more than six parameters the model that gives the best fit to the Table 1 data is $M(1)M(2)$, $M(3)$, $M(4)$. This model, involving dependence for outcomes of comparisons for given name and middle initial, does not fit particularly well. The likelihood ratio test statistic for lack of fit is 57.95 – an extreme value relative to the chi-square reference distribution with 9 degrees of freedom. The latent variable model $G(1)G(2)$, $G(3)$, $G(4)$ was estimated for each synthetic data set using iterative scaling. This model fit the synthetic data sets somewhat better than the model $M(1)M(2)$, $M(3)$, $M(4)$ fit the true match data. The largest lack of fit test statistic among the fifty synthetic data sets was 25.03 and the model was rejected only ten times at the 5% level of significance.

Averages of classification error rate estimates obtained using the synthetic data sets and the corresponding Monte Carlo standard errors are reported in Table 2 for true non-matches and Table 3 for true matches. After multiplication by 99, the error rates for true non-matches represent numbers of misclassified true non-matches divided by numbers of true matches. Results are given for the method of moments and iterative scaling, as well as the iterative method with $\mu^0 = 0.0000625$, 0.00025 and 0.001. The biases in estimated error rates for true non-matches are generally small. The iterative method with $\mu^0 = 0.001$

Table 2
Classification Error Rates, True Non-matches, Synthetic Data
(Monte Carlo Standard Errors in Parentheses)

Estimated Rate ($\times 99$)	Actual Rate ($\times 99$)				
	Method of Moments	Iter. Method $\mu^0 = 0.0000625$	Iter. Method $\mu^0 = 0.00025$	Iter. Method $\mu^0 = 0.001$	Iter. Scaling
0.02	0.0188 (0.0008)	0.0208 (0.0008)	0.0208 (0.001)	0.0207 (0.001)	0.0195 (0.001)
0.04	0.0381 (0.001)	0.0408 (0.0013)	0.0407 (0.0016)	0.0405 (0.0016)	0.0397 (0.0016)
0.06	0.057 (0.0012)	0.0626 (0.0015)	0.0615 (0.0018)	0.0602 (0.0019)	0.059 (0.0018)
0.08	0.076 (0.0015)	0.0855 (0.0017)	0.0838 (0.0019)	0.0804 (0.0022)	0.0785 (0.0019)
0.10	0.095 (0.0019)	0.1086 (0.0021)	0.1061 (0.0022)	0.1007 (0.0026)	0.0978 (0.0021)

Table 3
Classification Error Rates, True Matches, Synthetic Data
(Monte Carlo Standard Errors in Parentheses)

Estimated Rate	Actual Rate				
	Method of Moments	Iter. Method $\mu^0 = 0.0000625$	Iter. Method $\mu^0 = 0.00025$	Iter. Method $\mu^0 = 0.001$	Iter. Scaling
0.02	0.0580 (0.0013)	0.1179 (0.0041)	0.0507 (0.0014)	0.0149 (0.0008)	0.025 (0.0012)
0.04	0.0773 (0.0014)	0.1362 (0.004)	0.0735 (0.0012)	0.0359 (0.0018)	0.0455 (0.0016)
0.06	0.0966 (0.0014)	0.1542 (0.0038)	0.0954 (0.0012)	0.0660 (0.0014)	0.0646 (0.0018)
0.08	0.1159 (0.0014)	0.1722 (0.0036)	0.1165 (0.0012)	0.0866 (0.0017)	0.0841 (0.0019)
0.10	0.1348 (0.0014)	0.1904 (0.0035)	0.1319 (0.0014)	0.1025 (0.002)	0.1043 (0.002)

provides the best estimates, followed by iterative scaling. For true matches the performance of the iterative method is very sensitive to the choice of μ^0 . Although the iterative method performs well for $\mu^0 = 0.001$, the biases for $\mu^0 = 0.0000625$ and $\mu^0 = 0.00025$ are substantial. Estimates of classification error rates for true matches obtained using the method of moments also include large biases. Biases in estimates based on iterative scaling are relatively small.

Table 4

Classification Error Rates, True Non-matches,
Modified Synthetic Data
(Monte Carlo Standard Errors in Parentheses)

Estimated Rate ($\times 99$)	Actual Rate ($\times 99$)	
	Method of Moments	Iter. Scaling
0.02	0.0189 (0.0008)	0.0194 (0.001)
0.04	0.0385 (0.0011)	0.0396 (0.0016)
0.06	0.0577 (0.0013)	0.0589 (0.0019)
0.08	0.0767 (0.0016)	0.0785 (0.002)
0.10	0.0957 (0.002)	0.0978 (0.0021)

The information in Tables 4 and 5 is based on a series of synthetic data sets generated using a modified version of Table 1. Expected values of Table 1 cell counts under the model $M(1)M(2)$, $M(3)$, $M(4)$ were used for data generation. The biases in model-based classification error

rate estimates obtained using the method of moments are greatly reduced using the latent variable model $G(1)G(2)$, $G(3)$, $G(4)$ estimated using iterative scaling, particularly for true matches.

Table 5

Classification Error Rates, True Matches,
Modified Synthetic Data
(Monte Carlo Standard Errors in Parentheses)

Estimated Rate	Actual Rate	
	Method of Moments	Iter. Scaling
0.02	0.0553 (0.0014)	0.0208 (0.0011)
0.04	0.0747 (0.0014)	0.0415 (0.0016)
0.06	0.094 (0.0014)	0.0608 (0.0018)
0.08	0.1134 (0.0014)	0.0805 (0.002)
0.10	0.1325 (0.0015)	0.1007 (0.002)

6. COMPARISON OF ESTIMATION METHODS - REAL DATA

Results of comparisons of the three estimation methods using data from a record linkage application are presented in this section. Two data files used in empirical work reported by Fair and Lalonde (1987) were employed. The first file contained information on Ontario miners obtained from the Workmen's Compensation Board. The second file included information from the Canadian

Mortality Data Base (CMDB) for individual deaths during the period 1964 to 1977 inclusive. The miners' file included only those records with a valid social insurance number. The second file contained records that had survived an initial comparison exercise designed to eliminate records with no similarity to any of the records on the miners' file. The vital status of each miner at the end of 1977 had been classified as "confirmed dead", "confirmed alive" or "lost to follow-up" based on a previous linkage, combined with thorough follow-up procedures, including manual review. Records on the miners' file for individuals "confirmed dead" included the CMDB death registration number. More information on the construction of the files and the procedures used to determine true link status can be found in Fair and Lalonde.

Four identifiers – given name, NYSIIS code of mother's maiden name, day of birth and birth month – were chosen as matching fields for the comparison. Records on the miners' file with vital status "lost to follow-up" were eliminated. After records with missing values for at least one matching field or for birth year were also removed, file A (based on the miners' file) contained 45,638 records and file B (based on the CMDB) included 24,597 records. Restricting comparisons of the two files to pairs of records with the same NYSIIS representation of family name and birth years differing by at most one, there were 26,500 true non-matches and 2063 true matches.

Frequencies of outcomes among true matches and true non-matches are shown in Table 6. All loglinear models corresponding to a non-saturated latent variable model (that is, all models with fewer than eight parameters) are rejected by the frequency data for true non-matches at a very low level of significance. Among models with fewer than eight parameters the model $U(1)$, $U(2)U(4)$, $U(3)U(4)$ corresponds to the lowest likelihood ratio test statistic for lack of fit – 35.29. The model $M(1)$, $M(2)M(4)$, $M(3)M(4)$ provides an adequate fit to the true match data (likelihood ratio test statistic of 10.29).

Agreement probability estimates were computed using the method of moments, the iterative method and iterative scaling using the latent variable model $G(1)$, $G(2)G(4)$, $G(3)G(4)$. The likelihood ratio test statistic for the independence model corresponding to the method of moments estimator is 108 (six degrees of freedom). The independence model is rejected by the data at a very low significance level. In contrast, the likelihood ratio test statistic for the latent variable model $G(1)$, $G(2)G(4)$, $G(3)G(4)$ is 1.44 (two degrees of freedom), suggesting an adequate fit. Model-based estimates of classification error rates corresponding to each set of probability estimates were calculated for various thresholds. Actual classification error rates are compared to model-based estimates for true non-matches in Table 7 and true matches in Table 8. The error rates for true non-matches have been rescaled so that the number of true matches is in the denominator.

Table 6
Outcome Frequencies, Real Data

Outcome by Identifier: 0 = Disagreement, 1 = Agreement				Count	
Given Name	NYSIIS of Mother's Maiden Name	Day of Birth	Birth Month	True Matches	True Non- Matches
0	0	0	0	4	22,100
0	0	0	1	3	888
0	0	1	0	11	2,322
0	0	1	1	128	211
0	1	0	0	3	199
0	1	0	1	7	19
0	1	1	0	27	27
0	1	1	1	242	13
1	0	0	0	9	576
1	0	0	1	10	32
1	0	1	0	52	94
1	0	1	1	392	4
1	1	0	0	27	13
1	1	0	1	32	1
1	1	1	0	115	0
1	1	1	1	1,001	1
Total				2,063	26,500

Model-based classification error rate estimates obtained using the iterative method are very inaccurate, particularly for true non-matches, regardless of the value of μ^0 . Error rate estimates obtained using iterative scaling are slightly less accurate than estimates based on the method of moments for true matches. However, they are considerably more accurate than method of moments estimates for true non-matches.

Some words of caution are necessary. Even though the model $U(1)$, $U(2)U(4)$, $U(3)U(4)$ does not adequately describe the dependencies among true non-matches, the iterative scaling algorithm obtained a good fit using an estimate of the proportion of matched records (0.0747) that differs somewhat from the true value (0.0722). A similar fit can also be obtained using the model $G(1)G(2)$, $G(1)G(3)$, $G(4)$ and an estimate of 0.077 for the proportion of matches. Error rate estimates based on the model $G(1)G(2)$, $G(1)G(3)$, $G(4)$ are no better than estimates obtained using the method of moments.

Table 7
Classification Error Rates, True Non-matches, Real Data

Estimated Rate ($\times 12.84$)	Actual Rate ($\times 12.84$)				
	Method of Moments	Iter. Method $\mu^0 = 0.0000625$	Iter. Method $\mu^0 = 0.00025$	Iter. Method $\mu^0 = 0.001$	Iter. Scaling
0.02	0.0368	1.311	0.1859	0.186	0.0339
0.04	0.0796	1.314	0.1888	0.193	0.0649
0.06	0.1224	1.317	0.1917	0.1967	0.0684
0.08	0.1573	1.323	0.1990	0.1994	0.1106
0.10	0.1863	1.333	0.60	0.4066	0.1282

Table 8
Classification Error Rates, True Matches, Real Data

Estimated Rate	Actual Rate				
	Method of Moments	Iter. Method $\mu^0 = 0.0000625$	Iter. Method $\mu^0 = 0.00025$	Iter. Method $\mu^0 = 0.001$	Iter. Scaling
0.02	0.0166	0.0141	0.0193	0.0225	0.0105
0.04	0.0318	0.0264	0.029	0.0278	0.0263
0.06	0.0598	0.0383	0.0472	0.0326	0.0529
0.08	0.0782	0.0416	0.1372	0.0488	0.0784
0.10	0.0966	0.045	0.1393	0.1371	0.0958

7. CONCLUSIONS

In this paper, the issue of classification error rate estimation for record linkage has been discussed. The Fellegi-Sunter framework provides for the calculation of classification error rate estimates using estimates of agreement probabilities. These model-based estimates typically have poor properties in practice. It has been demonstrated that their properties can be improved through careful estimation of agreement probabilities. Three estimation methods have been evaluated using synthetic data as well as information from a real application.

For two of the three methods, the assumption that outcomes of comparisons for different data fields are independent was employed. This assumption was not valid for either the synthetic data or the real data. The synthetic data included strong dependencies for true matches and minor dependencies for true non-matches. Dependencies in the real data were particularly strong for true non-matches. Classification error rate estimates obtained using the method of moments, which relies on the assumption of independence, included substantial bias for synthetic data and were relatively inaccurate for real data. The magnitude of the bias in classification error rate estimates for synthetic data obtained using the iterative method

depended on the definition of an initial set of matches. Although some definitions of the initial set of matches led to relatively small biases, others produced estimates with biases much larger than those obtained using the alternative methods. For the real data, all the definitions of the initial set of matches considered led to very inaccurate error rate estimates. There are no mathematical rules available for the choice of an initial set of matches for the iterative method. The results in this paper provide no evidence to recommend its use.

The third method relies on a parameterization of dependencies between outcomes of comparisons for different data fields using loglinear effects. Under this parameterization, estimates of agreement probabilities that do not rely on the independence assumption can be obtained through use of the iterative scaling method to estimate the parameters of a latent variable model. For the synthetic data sets with lack of independence, model-based classification error rate estimates obtained using iterative scaling included much smaller biases than estimates based on the independence assumption. Although the latent variable model fit most synthetic data sets better than a model based on the independence assumption, it sometimes exhibited significant lack of fit. When the synthetic data was modified to improve the fit of the latent variable

model, there was no evidence of bias in model-based classification error rate estimates. The real data included important departures from independence for both true matches and true non-matches. Model-based error rate estimates obtained using iterative scaling were slightly less accurate than estimates based on the method of moments for true matches and considerably more accurate for true non-matches.

The results reported here indicate that properties of model-based classification error rates estimates can be improved using an appropriate estimator of agreement probabilities. Latent variable models and iterative scaling provide a method of incorporating dependencies between outcomes of comparisons for different data fields during estimation of agreement probabilities.

ACKNOWLEDGEMENTS

The authors would like to thank William Winkler for providing the computer code that was the basis of the iterative scaling estimation program used to obtain our results, as well as Fritz Scheuren and three anonymous referees for comments on a earlier version of this paper that led to a significant improvement in both the content and the presentation. Thanks are also due to Martha Fair and Pierre Lalonde for making available the Ontario miners' data and the outcome frequency data for true matches.

REFERENCES

- BARTLETT, S., KREWSKI, D., WANG, Y., and ZIELINSKI, J.M. (1993). Evaluation of error rates in large scale computerized record linkage studies. *Survey Methodology*, 19, 3-12.
- BELIN, T.R. (1990). A proposed improvement in computer matching techniques. In *Statistics of Income and Related Administrative Record Research: 1988-1989*, U.S. Internal Revenue Service, 167-172.
- BELIN, T.R., and RUBIN, D.B. (1991). Recent developments in calibrating error rates for computer matching. *Proceedings of the Annual Research Conference, U.S. Bureau of the Census*, 657-668.
- FAIR, M.E., and LALONDE, P. (1987). Missing identifiers and the accuracy of individual follow-up. *Proceedings: Symposium on Statistical Uses of Administrative Data, Statistics Canada*, 95-107.
- FAIR, M.E., NEWCOMBE, H.B., and LALONDE, P. (1988). Improved mortality searches for Ontario miners using social insurance index identifiers. Research report, Atomic Energy Control Board.
- FELLEGI, I.P., and SUNTER, A.B. (1969). A theory for record linkage. *Journal of the American Statistical Association*, 64, 1183-1210.
- HABERMAN, S.J. (1976). Iterative scaling procedures for log-linear models for frequency tables derived by indirect observation. *Proceedings of the Statistical Computing Section, American Statistical Association*, 45-50.
- HABERMAN, S.J. (1979). *Analysis of Qualitative Data*. London: Academic Press.
- IMSL (1987). Math/Library FORTRAN subroutines for mathematical applications. Houston: IMSL Inc.
- MORÉ, J., GARROW, B., and HILLSTROM, K. (1980). User guide for MINPACK-1. Argonne National Labs Report ANL-80-74.
- NEWCOMBE, H.B., KENNEDY, J.M., AXFORD, S.J., and JAMES, A.P. (1959). Automatic linkage of vital records. *Science*, 130, 954-959.
- NEWCOMBE, H.B. (1988). *Handbook of Record Linkage: Methods for Health and Statistical Studies, Administration, and Business*. Oxford: Oxford University Press.
- THIBAUDEAU, Y. (1989). Fitting log-linear models in computer matching. *Proceedings of the Statistical Computing Section, American Statistical Association*, 283-288.
- WINKLER, W.E. (1989). Near automatic weight computation in the Fellegi-Sunter model of record linkage. *Proceedings of the Annual Research Conference, U.S. Bureau of the Census*, 145-155.

Robust Joint Modelling of Labour Force Series of Small Areas

D. PFEFFERMANN and S.R. BLEUER¹

ABSTRACT

In this article we report the results of fitting a state-space model to Canadian unemployment rates. The model assumes an additive decomposition of the population values into a trend, seasonal and irregular component and separate autoregressive relationships for the six survey error series corresponding to the six monthly panel estimators. The model includes rotation group effects and permits the design variances of the survey errors to change over time. The model is fitted at the small area level but it accounts for correlations between the component series of different areas. The robustness of estimators obtained under the model is achieved by imposing the constraint that the monthly aggregate model based estimators in a group of small areas for which the total sample size is sufficiently large coincide with the corresponding direct survey estimators. The performance of the model when fitted to the Atlantic provinces is assessed by a variety of diagnostic statistics and residual plots and by comparisons with estimators in current use.

KEY WORDS: Design variance; Kalman filter; Panel survey; Rotation bias; State-space model.

1. INTRODUCTION

A time series model for survey data is the combination of two distinct models. The “census model” describing the evolution of the finite population values over time and the survey errors model representing the time series relationships between the survey errors of the survey estimators. There are at least four main reasons for wishing to model the raw survey estimators:

- (a) The model based estimators of the population values resulting from the modelling process have in general smaller variances than the survey estimators, particularly in small areas where the sample sizes are small.
- (b) The model we employ yields estimators for the seasonal effects and for the variances of these estimators as a by-product of the estimation process.
- (c) The model can be used to forecast the population values, the trend and the seasonal components for time periods beyond the sample time period for which the direct survey estimators are available. Such forecasts are important when assessing the performance of the model and for policy decision making.
- (d) The model can be used to detect turning points in the level of the series and assess their significance. (Work on this problem will be addressed in a separate article).

The methodology described in this article integrates the methodologies presented in Pfeffermann and Burck (1990) and Pfeffermann (1991) with some new modifications and extensions. The main features of the model are as follows:

1. The model decomposes the population values into the unobservable components of trend, seasonality and irregular terms. Smoothed predictors of the three

components (and hence of the population values) based on all the available data, and standard errors of the prediction errors are obtained straightforwardly by application of the Kalman filter. The standard errors are modified to account for the extra variation induced by the use of estimated parameter values.

2. The model uses the distinct monthly panel estimators as input data. The use of the panel estimators has two important advantages over the use of the mean estimators: (i) It identifies better the time series model holding for the survey errors by analysing contrasts between the panel estimators, (ii) It yields more efficient estimators for the model parameters and hence better predictors for the unobservable model components.
3. The model accounts for changes in the variances of the survey errors over time and for possible rotation group effects.
4. The model can be applied simultaneously to the panel estimators in separate small areas. The census model is extended in this case to account for the cross-correlations between the unobservable components of the population values operating in these areas.
5. A modification to ensure the robustness of the small area estimators against possible model breakdowns is incorporated into the model equations. The modification consists of constraining the model based estimators of aggregates of the population values over a group of small areas for which the total sample size is sufficiently large to coincide with the corresponding aggregate survey estimators. As a result, sudden changes in the level of the series are reflected in the model based estimators with no time lag.

¹ D. Pfeffermann, Department of Statistics, Hebrew University, Jerusalem 91905; S.R. Bleuer, Social Surveys Methods Division, Statistics Canada, Ottawa, Ontario, K1A 0T6.

The model and the robustness modifications are described in more detail in section 2. Empirical results obtained when fitting the model to the four Atlantic provinces of Canada are presented in section 3. Section 4 contains a short summary with suggestions for extension of the analysis.

Before concluding this section we mention that in the U.S., the state unemployment estimates are produced for most of the states based on time series models which have a similar structure to the model used in our study. See Tiller (1992) for details. A major difference between the two models is that in the U.S., the model postulated for the population values includes also explanatory variables so that the trend and the seasonal component only account for the trend and seasonal variations not accounted for the explanatory variables. The models fitted to the survey errors are like in our case of the ARIMA type and they likewise account for changes in the variances of the survey errors. They are otherwise different because of the very different sample rotation schemes used in the two countries. Another notable difference between the two models is that in the U.S., the models are fitted to each state separately and the input data consist of only the mean survey estimates, that is, one observation for every month. As a result, the models do not account for rotation group biases.

2. A STATE-SPACE MODEL FOR CANADA UNEMPLOYMENT SERIES

2.1 The Canadian Labour Force Survey

Data on unemployment are collected as part of the Labour Force Survey (LFS) carried out by Statistics Canada. The Canadian LFS is a rotating monthly panel survey by which every new sampled panel of households is retained in the sample for six successive months before being replaced by another panel from the same PSU's or strata. The PSU's are defined by geographic locations (city blocks or urban centers in the urban regions and groups of enumeration areas in the rural regions). The strata are homogeneous groups of PSU's defined by geographic locations such as city tracts, census subdivisions and enumeration areas. In the urban regions, (about 2/3 of the sample), every PSU is represented in only one panel. In the rural regions, the PSU's are represented in all the panels but with different enumeration areas in different panels. As a result, the separate panel estimators can be assumed to be independent, a property validated and utilized in other studies, see *e.g.* Lee (1990). For a recent report describing the design of the LFS and the construction of the direct survey estimators, the reader is referred to Singh *et al.* (1990).

2.2 The Census Model

In what follows we consider a single small area. In section 2.4 we consider joint modelling of the panel estimates in a group of small areas. The model postulated for the population values is the Basic Structural Model (BSM) which consists of the following set of equations.

$$Y_t = L_t + S_t + \epsilon_t; \quad L_t = L_{t-1} + R_{t-1} + \eta_{Lt};$$

$$R_t = R_{t-1} + \eta_{Rt}; \quad \sum_{j=0}^{11} S_{t+j} = \eta_{St}. \quad (2.1)$$

In (2.1) Y_t is the population value ("true" unemployment rate) at time t , L_t is the trend level, R_t is the increment, S_t the seasonal effect and ϵ_t the irregular term assumed to be white noise with zero mean and variance σ_ϵ^2 . Thus, the first equation in (2.1) postulates the classical decomposition of a time series into a trend, seasonal and irregular components. This decomposition is inherent in the commonly used procedures for seasonal adjustment, see *e.g.* Dagum (1980). Notice however that in the present case the series $\{Y_t\}$ is itself unobservable. The series $\{\eta_{Lt}\}$, $\{\eta_{Rt}\}$ and $\{\eta_{St}\}$ are independent white noise disturbances with mean zero and variances σ_L^2 , σ_R^2 and $\sigma_S^2 \times g(t)$ respectively. Hence, the second and third equations of (2.1) define a local approximation to a linear trend whereas the last equation models the evolution of the seasonal effects such that the sum of every 12 successive effects fluctuates around zero. Notice that the variances of the error terms η_{St} are time dependent. The functions $g(t)$ are specified at the end of section 3.1.

The theoretical properties of the BSM in comparison to other models are discussed in Harrison and Stevens (1976), Harvey (1984) and Maravall (1985). Empirical results illustrating the performance of the model are shown in Harvey and Todd (1983), Morris and Pfeffermann (1984) and Pfeffermann (1991). Although more restricted than the family of ARIMA models, the BSM is now recognized as being flexible enough to approximate the behaviour of many diverse time series.

2.3 The Survey Errors Model

The model holding for the survey errors was identified initially by analyzing separately the pseudo error series $e_{t,p}^{(j)} = (y_t^{(j)} - \bar{y}_t)$, $t = 1, \dots, N$, where $y_t^{(j)}$ is the estimator of Y_t based on j -th panel $j = 1, \dots, 6$, (the panel surveyed for the j -th successive month) and $\bar{y}_t = \sum_{j=1}^6 y_t^{(j)} / 6$ is the mean estimator. Notice that $(y_t^{(j)} - \bar{y}_t) = (e_t^{(j)} - \sum_{j=1}^6 e_t^{(j)} / 6)$, where $e_t^{(j)} = (y_t^{(j)} - Y_t)$ are the true survey errors. Thus, the notable feature of the contrasts $(y_t^{(j)} - \bar{y}_t)$ is that they are functions of only the survey errors irrespective of the model holding for the population values.

There are two prior considerations in the choice of a model for the survey errors:

- (a) The model should account for possible rotation group biases or more generally, allow for different means for the survey errors of different panels.
- (b) The model should account for changes in the variances of the survey errors over time.

Rotation group biases may arise from providing different information on different rounds of interview, depending on the length of time that respondents are included in the sample, or on the method of data collection, say, whether by telephone or by home interview. (In the Canadian LFS, the first panel is interviewed by home visits, the other panels are interviewed by telephone). Another possible reason for differences between the panel survey error means is differences in the nonresponse patterns across the panels. See Pfeffermann (1991) for further discussion with references to earlier studies on this problem.

Changes in the variances of the survey errors over time occur when the variances are function of the level of the series. Indeed, as revealed by figure 1 in section 3, the estimates of the standard deviations of the survey errors are subject to seasonal effects with a seasonal pattern that follows the seasonal pattern of the population values. Another possible explanation for changes in the variances of the survey errors is changes in the sampling design. For example, the overall sample size of the Canadian LFS was reduced in 1985-1986 from 55,000 households to 48,000 households. This reduction in the sample size was associated with other changes in the design. See Singh *et al.* (1990) for details.

Application of simple model estimation and diagnostic procedures to the pseudo survey errors suggest a 3rd order autoregressive (AR) model for the standardized survey errors $\tilde{e}_t^{(j)} = (e_t^{(j)} - \beta_j)/SD(e_t^{(j)})$, i.e.

$$\tilde{e}_t^{(j)} = \phi_{j1} \tilde{e}_{t-1}^{(j-1)} + \phi_{j2} \tilde{e}_{t-2}^{(j-2)} + \phi_{j3} \tilde{e}_{t-3}^{(j-3)} + u_t^{(j)}, j = 1, \dots, 6, \quad (2.2)$$

where $\beta_j = E(e_t^{(j)})$ are the rotation group biases, $SD(e_t^{(j)})$ are the design standard deviations and $u_t^{(j)}$ are independent white noise with mean zero and variances σ_j^2 . It is assumed that $\sum_{j=1}^6 \beta_j = 0$ which implies that the mean survey estimator, \bar{y}_t , is unbiased. See Pfeffermann (1991) for discussion on the need to constraint the bias coefficients. Subsequent analysis when fitting the combined model defined by (2.1) and (2.2) (see section 2.4) validates this model with the further observation that the coefficients ($\phi_{j1}, \phi_{j2}, \phi_{j3}$) can be assumed to be equal for $j = 4, 5, 6$. Furthermore, for the first panel an AR(1) model already gives a good fit whereas for the second and third panel an AR(2) model is appropriate although with different

coefficients. These relationships hold for each of the four Atlantic provinces.

One of the referees of this article raised the question of whether the AR(3) model defined by (2.2) is flexible enough to account for the panel estimates correlations at high lags which are believed to be high because of "PSU effects". As mentioned in section 2.1, panels rotating out of the sample are replaced by panels from the same PSU's and it usually takes several years before a PSU is exhausted and replaced by a neighbouring PSU. Lee (1990) presents two sets of panel estimates correlations for the Canadian LFS. The first set, denoted by ρ_j , are the correlations between estimates produced from the same panel so that j ranges from 1 to 5. The second set, denoted by γ_j , are the correlations between estimates produced from a panel and its predecessor so that j ranges from 1 to 11. The ρ -correlations are generally high as expected but it should be emphasized that they are lower for the unemployment series than for the employment series, demonstrating the high mobility of the unemployment Labour Force. The γ -correlations are much smaller than the ρ correlations but as mentioned by the author, the computation of these correlations is much less reliable and their behavior is somewhat fuzzy showing occasionally an increasing trend. We computed the serial correlations based on the models (2.2) with the ϕ -coefficients replaced by their estimated values and found in general a close fit to the ρ -correlations at all the lags from 1 to 5. The correlations at higher lags are different from the corresponding γ -correlations but interesting enough, they are in most cases higher and always decrease as j increases.

Another question related to the model (2.2) raised by the referees is whether one could apply the log transformation to the raw data for stabilizing the survey error variances, rather than modelling the standardized errors. There are two main reasons for not using the log transformation in our case. Foremost, the use of this transformation would imply a multiplicative decomposition for the population unemployment rates which is counter to common practice of postulating an additive decomposition. In Statistics Canada the unemployment rates in the two larger provinces out of the four considered in our study are deseasonalized by postulating the additive decomposition. In the U.S. the models fitted to the state unemployment series likewise postulate an additive decomposition. See Tiller (1992). The second reason is that changes in the survey error variances may result from changes in the sampling design and in particular, from changes in the sample sizes. Such changes cause discrete shifts in the variances which cannot be handled effectively by the log transformation. As noted also by one of the referees, transforming the data has the drawback of producing nonlinearity in aggregating the estimates over the panels and/or the small areas.

The model defined by (2.2) satisfies the two prior considerations discussed above. The actual application of the model requires however two modifications:

1. For the first three panels there is not a long enough history to permit the fitting of an AR(3) model. For example, the survey error $e_t^{(1)}$ corresponds to the panel which is in the sample for the first time. In order to overcome this problem, we replace the missing survey errors by the survey errors corresponding to the panels previously selected from the same PSU's or strata. For example, the AR(2) model fitted to $\tilde{e}_t^{(2)}$ is

$$\tilde{e}_t^{(2)} = \phi_{21} \tilde{e}_{t-1}^{(1)} + \phi_{22} \tilde{e}_{t-2}^{(6)} + u_t^{(2)}. \quad (2.3)$$

Notice that the panel surveyed for the second time at month t replaces at time $(t - 1)$ the panel surveyed for the sixth time at month $(t - 2)$ so that both panels represent the same PSU's or strata. The use of surrogate survey errors in the case of the first three panels may explain the different models identified for these panels as compared to the model identified for the other three panels.

2. The true standard deviations of the survey errors are unknown whereas the survey estimates of the standard deviations are themselves subject to sampling errors. To overcome this problem, we use smoothed values of the estimated standard deviations, obtained by fitting the relationship

$$(\tilde{SD})_t = \hat{\gamma} (\hat{SD})_{t-1} + \hat{\gamma}_0 t + \sum_{i=1}^{12} \hat{\gamma}_i D_{it}, \quad (2.4)$$

with the γ -coefficients estimated by ordinary least squares. The notation $(\hat{SD})_t$ defines the raw, unsmoothed estimate of the design standard deviation of the mean survey estimator, \bar{y}_t , at month t and $\{D_{it}\}$ are dummy variables accounting for monthly seasonal effects so that $D_{it} = 1$ when $t = 12k + i$, $k = 0, 1, \dots$, $i = 1, \dots, 12$ and $D_{it} = 0$ otherwise. The smoothed standard deviations of the panel survey errors are obtained as $\tilde{SD}(e_t^{(j)}) = \sqrt{6}(\tilde{SD})_t$. The latter estimates are used as surrogates for the true, unknown, standard deviations.

2.4 State-space Representation and Estimation of the Model Holding for the Survey Estimators

It follows from (2.1) that the panel estimators can be modeled as

$$y_t^{(j)} = L_t + S_t + \epsilon_t + e_t^{(j)}, \quad j = 1, \dots, 6, \quad (2.5)$$

where

$$L_t = L_{t-1} + R_{t-1} + \eta_{Lt}; \quad R_t = R_{t-1} + \eta_{Rt};$$

$$\sum_{j=0}^{11} S_{t+j} = \eta_{St}, \quad (2.6)$$

with $\{\epsilon_t\}$, $\{\eta_{Lt}\}$, $\{\eta_{Rt}\}$ and $\{\eta_{St}\}$ defined as in (2.1). The separate models defined by (2.5), (2.6) and (2.2) can be cast into a compact state-space representation with $y_t' = (y_t^{(1)}, \dots, y_t^{(6)})$ as the input data, similar to the representation in Pfeffermann (1991). Following that representation, the survey errors (and in the present study also the census irregular terms) are included as part of the state vector so that there are no residual terms in the observation equation defined by (2.5). Unlike in Pfeffermann (1991), however, the transition matrix and the Variance-Covariance (V-C) matrix of the state error terms are not fixed in time since they depend on the design variances of the survey errors which, as explained in section 2.3, change over time.

The state-space representation of the model permits us to update, smooth or predict the state vectors and hence the seasonal, trend and population values at any given month t by means of the Kalman filter. Denote by α_t the state vector corresponding to month t . The state vector comprises the trend level, increment and seasonal effects, the rotation group biases and the survey errors. See Pfeffermann (1991) for details. By “updating” we mean estimation of α_t at month t based on all the data until and including month t . “Smoothing” refers to the estimation of α_t based on all the available data for all the months before and after month t . Smoothing is required for improving past estimates as, for example, when estimating the seasonal effects or when estimating changes in the population values or the trend levels. “Prediction” of state vectors corresponding to postsample months is important for policy making. Predictions within the sample period allow to assess the performance of the model, e.g. by comparing the forecasted panel estimates as derived from the predicted state vectors with the actual estimates. See section 3 for details. The theory of state-space models and the Kalman filter is developed in numerous publications, see Pfeffermann (1991) for the filtering and smoothing equations with references. Notice that the filtering and the smoothing equations not only yield the three sets of estimators for any given month t but also the V-C matrices of the corresponding estimation errors.

The actual application of the Kalman filter requires the estimation of the unknown model parameters and the initialization of the filter, that is, the estimation of the initial state vector α_0 and the corresponding V-C matrix of the estimation errors. For a single small area, the unknown model parameters are the four variances of the error terms in the census model (2.1) and the eight

autoregression coefficients and six residual variances in the panel survey error models (2.2). (The rotation group means are included in the state vectors as fixed, time invariant coefficients). In order to reduce the number of free parameters in the combined state-space model, we assume $\sigma_j^2 = \sigma^2 \times \tilde{\sigma}_j^2, j = 1, \dots, 6$, where $\{\sigma_j^2\}$ are the residual variances in (2.2) and $\tilde{\sigma}_j^2$ are the estimates of the residual variances obtained by fitting the autoregression equations to the pseudo survey errors $e_{t,p}^{(j)}$ defined in section 2.3. This assumption reduces the number of unknown parameters from 18 to 13. (The estimates $\tilde{\sigma}_j^2$ are very close for $j = 4, 5, 6$ and have been set equal).

Assuming that the error terms in the census and survey error models have a normal distribution, the unknown model parameters can be estimated by maximization of the likelihood. See Pfeiffermann and Burck (1991) for a brief description of the application of the method of scoring maximization algorithm and for the initialization of the filter. That article includes references to more rigorous discussions.

2.5 Adjustments to Account for the Use of Estimated Parameter Values

Once the unknown model parameters have been estimated, the Kalman filter equations can be applied with the true parameter values replaced by the parameter estimates. As noted in section 2.4, the Kalman filter not only produces estimates for the state vectors but also the V-C matrices of the corresponding estimation errors. A possible problem arising from the use of these V-C matrices, however, is that they ignore the extra variation implied by parameter estimation, thus resulting in underestimation of the true variances.

Formally, let $\hat{q}_t(\hat{\lambda})$ define the estimator of q_t at month t , based on all the data available until some given month n , where $\hat{\lambda}$ represents the estimators of the unknown model parameters. The estimation error can be decomposed as

$$[\hat{q}_t(\hat{\lambda}) - \alpha_t] = [\hat{q}_t(\lambda) - \alpha_t] + [\hat{q}_t(\hat{\lambda}) - \hat{q}_t(\lambda)], \quad (2.7)$$

which is the sum of the error if λ were known plus the error due to estimation of λ . The two terms in the right-hand side of (2.7) are uncorrelated. A simple way to verify this property is by noting that $\hat{q}_t(\lambda) = E(q_t | Y, \lambda)$ where Y represents all the available data. By conditioning on Y and λ , $[\hat{q}_t(\hat{\lambda}) - \hat{q}_t(\lambda)]$ is nonstochastic whereas $E\{[\hat{q}_t(\lambda) - \alpha_t] | Y, \lambda\} = 0$. It follows therefore from (2.7) that

$$\begin{aligned} Q_t &= E\{[\hat{q}_t(\hat{\lambda}) - \alpha_t][\hat{q}_t(\hat{\lambda}) - \alpha_t]'\} \\ &= E\{[\hat{q}_t(\lambda) - \alpha_t][\hat{q}_t(\lambda) - \alpha_t]'\} \\ &\quad + E\{[\hat{q}_t(\hat{\lambda}) - \hat{q}_t(\lambda)][\hat{q}_t(\hat{\lambda}) - \hat{q}_t(\lambda)]'\} \\ &= A_t + B_t. \end{aligned} \quad (2.8)$$

In order to estimate A_t and B_t we condition on Y and follow the approach proposed by Hamilton (1986). By this approach, realizations $\lambda_{(k)}, k = 1, \dots, K$ are generated from the asymptotic normal posterior distribution of λ , that is, from a $N(\hat{\lambda}, \hat{\Lambda})$ distribution where $\hat{\lambda}$ is the maximum likelihood estimator of λ and $\hat{\Lambda}$ is the asymptotic V-C matrix of $\hat{\lambda}$. (Both $\hat{\lambda}$ and $\hat{\Lambda}$ are obtained from the method of scoring). The Kalman filter is then applied with each of these realizations yielding estimates $\hat{q}_t(\lambda_{(k)})$ with V-C matrices $P_t(\lambda_{(k)})$. The matrices A_t and B_t are estimated as

$$\begin{aligned} \hat{A}_t &= \frac{1}{K} \sum_{k=1}^K P_t(\lambda_{(k)}); \\ \hat{B}_t &= \frac{1}{K} \sum_{k=1}^K [\hat{q}_t(\lambda_{(k)}) - \hat{q}_t(\hat{\lambda})][\hat{q}_t(\lambda_{(k)}) - \hat{q}_t(\hat{\lambda})]'. \end{aligned} \quad (2.9)$$

Ansley and Kohn (1986) propose an estimator for B_t based on first order Taylor series approximation. The use of their estimator is computationally less intensive but the procedure proposed by Hamilton is somewhat more flexible in terms of the assumptions involved and it enables a better insight into the sensitivity of the Kalman filter output to errors in the parameter estimators.

2.6 Joint Modelling in Several Small Areas

The model considered so far refers to a single area. When the sample sizes in the various areas are small, more efficient estimators can often be derived by modelling in addition the cross-sectional relationships between the area population values. Clearly, the increase in efficiency resulting from such joint modelling depends on the sample sizes within the small areas and the closeness of the behaviours of the area population values over time.

The survey errors are independent between the areas so that any joint modelling of the survey estimators applies only to the census model. For modelling the unemployment rates in the four Atlantic provinces, we follow Pfeiffermann and Burck (1990) and allow for nonzero contemporary correlations between corresponding error terms of the census models operating in these provinces. Thus, if $y'_{t,a} = (\epsilon_t^{(a)}, \eta_{Lt}^{(a)}, \eta_{Rt}^{(a)}, \eta_{St}^{(a)})$ denotes the vector of error terms at time t associated with the census model operating in area a , it is assumed that $C_{a,b} = E(y_{ta} y'_{tb})$ is diagonal but with possibly non zero covariances on the main diagonal. The actual implication of this assumption is that if, for example, there is a significant increase in the trend level in one province, similar increases can be expected to occur in other provinces.

The resulting joint model holding for the four provinces (or more generally for a group of areas) can again be cast into a state-space form, see equations (2.7) and (2.8) in

Pfeffermann and Burck (1990). A major problem with the fitting of this model, however, is the joint estimation of all the unknown parameters which is computationally too intensive in terms of computer time and storage space. (The computer program written for the application of the method of scoring uses numerical first order derivatives so that each derivative requires a separate sweep through all the data. Each sweep involves the computation of the Kalman filter equations for each month included in the sample period).

To deal with this problem, we first fitted the models defined by (2.5), (2.6) and (2.2) separately for each of the provinces. We also postulated equal correlations between the corresponding error terms of the separate census models across the provinces so that

$$\phi_{a,b} = C_{a,a}^{-1/2} C_{a,b} C_{b,b}^{-1/2} = \phi \quad 1 \leq a, b \leq 4, \quad (2.10)$$

where $C_{a,a} = E(y_{ta} y'_{ta})$. The four correlations maximizing the likelihood of the joint model were determined by a grid search procedure with the other model parameters held fixed at their previously estimated values.

The assumption of equal correlations reduces the number of unknown parameters considerably. It can be justified also by the small number of areas considered for this study implying that no other pre-imposed structure on these correlations can be safely detected. More substantively, a simple breakdown of the Labour Force by industry (Table 1 of Section 3) shows very similar relative frequencies in the four provinces suggesting a high degree of homogeneity in their economies.

2.7 Modifications to Protect Against Model Failures

The use of a model for the production of official statistics raises the question of how to protect against possible model failures. As discussed below, testing the model every time that new data becomes available is not feasible requiring instead the development of a built-in mechanism to ensure the robustness of the estimators when the model fails to hold.

For modelling the Labour Force series in small areas we employed the modification proposed by Pfeffermann and Burck (1990). By this modification, the updated state vector estimates at any given time t , are constrained to satisfy the condition

$$\sum_{a=1}^A w_{ta} \hat{Y}_{ta} = \sum_{a=1}^A w_{ta} \bar{y}_{ta} \quad t = 1, 2, \dots, \quad (2.11)$$

where \hat{Y}_{ta} is the model based estimator of the population value Y_{ta} in area a , $\bar{y}_{ta} = 1/6 \sum_{j=1}^6 y_{ta}^{(j)}$ is the corresponding survey estimator and $w_{ta} = M_{ta}/M_t$ is the relative size of the Labour Force in that area so that $M_t = \sum_{a=1}^4 M_{ta}$ and $\sum_{a=1}^4 w_{ta} = 1$. Notice that $\sum_{a=1}^A w_{ta} \hat{Y}_{ta}$

and $\sum_{a=1}^A w_{ta} \bar{y}_{ta}$ are correspondingly the model based estimator and the direct survey estimator of the aggregate population value in the group of areas considered. The condition 2.11 can be written alternatively as $\sum_{a=1}^A w_{ta} \bar{e}_{ta} = 0$ where $\bar{e}_{ta} = \sum_{j=1}^6 e_{ta}^{(j)}/6$ is the mean survey error for state a . Pfeffermann and Burck (1990) show how to modify the Kalman filter equations so that it produces the constrained state vector estimator and its correct V-C matrix under the model (without the constraint), for every month t .

The rationale behind the modification is simple. It assumes that the total sample size in all the areas is sufficiently large and hence that the aggregate survey estimators can be trusted. This assumption in fact dictates the level of aggregation required, see below. By constraining the aggregate model based estimators to coincide with the aggregate survey estimators, the analyst ensures that any real change in the population values reflected in the survey estimators will be likewise reflected in the model based estimators. Notice that without constraining the estimators, sudden changes in the level of the series, for example, will be reflected in the model based estimators only after several months because these estimators depend not only on current data but also on past data. On the other hand, if no substantial changes occur, the model based estimators can be expected to satisfy approximately the constraints even without imposing them explicitly. Thus, the constrained estimators should perform almost as well as the unconstrained estimators in regular time periods.

The assumption that the total sample size in all the areas is large and hence that the aggregate survey estimator is sufficiently close to the corresponding population value is critical. It guarantees (in high probability) that the modification will only occur when there are real changes in the population values and not as a result of large sampling errors. Admittedly, and as noted by one of the referees, in the application of the method to the Atlantic provinces described in section 3, the aggregate estimator is based on only four provinces so that its standard error is about 50 percent of the standard errors of the province survey estimators, depending on the province sample sizes. (The province survey estimators are independent, conditional on the corresponding population province values). Thus, if the constraints are to be used in practice, the aggregation should be carried out over a larger set of provinces or other small areas.

The following two alternative approaches have been suggested for dealing with the robustness problem:

- (i) Perform a time series outlier detection as proposed for example in Chang, Tiao and Chen (1988).
- (ii) Model the time series of proportions $\{\hat{\pi}_{ta} = \bar{y}_{ta}/\sum_{a=1}^A \bar{y}_{ta}, a = 1, \dots, (A - 1)\}$ if these time series exhibit smoother behavior than the series $\{\bar{y}_{ta}\}$.

The detection of outliers is an important aspect of any modelling exercise but the question remaining is how to modify the population value estimates once observations (survey estimates) are detected as outliers. Notice in this respect that our main concern is with current estimates that is, the most recent available estimates. In Chang, Tiao and Chen (1988), the motivation for the outlier detections is to *remove* their effect from the observations so as to better understand the underlying structure of the series and improve the estimation of the model parameters. But if the cause of an outlier observation is a real shift in the level of the population values, this shift should not be removed but rather accounted for in the model based estimators. Harrison and Stevens (1976) propose to account for such changes by modifying the prior distribution of the state vectors, *e.g.* by increasing the variances of the state vector errors so as to allow for more rapid changes in the state vector estimators. See Morris and Pfeiffermann (1984) for an example. Our approach of constraining the model based estimators to coincide with aggregate survey estimators provides a more automatic procedure that does not require timely prior information.

The second approach suggested for dealing with the robustness problem is appealing since abrupt changes in the population values can be expected to cancel out in the ratios $\hat{\pi}_{ta}$. The main disadvantage of the use of this approach is that the model holding for the 'true' ratios π_{ta} is naturally very different from the model holding for the population values Y_t as defined by (2.1) and in particular, it no longer provides estimates for the trend and the seasonal effects which, as mentioned in the introduction, is one of the major uses of our approach. It is also not clear how to extract the estimates for the population values Y_t from the model holding for the ratios $\hat{\pi}_{ta}$, without some additional assumptions, like, for example, our assumption that the aggregate survey estimator is sufficiently close to the corresponding population value.

The use of constraints of the form (2.11) was previously considered by Battese, Harter and Fuller (1988) and by Pfeiffermann and Barnard (1991) for analyzing cross-sectional surveys. Pfeiffermann and Burck (1990) present empirical results illustrating the good performance of the modified estimators in abnormal time periods. See also section 3.

3. FITTING THE MODEL TO THE ATLANTIC PROVINCES, EMPIRICAL RESULTS

The model defined by (2.2), (2.5) (2.6) and (2.10) was fitted to the monthly panel estimators in the four Atlantic provinces in two stages. In the first stage the model defined by (2.2), (2.5) and (2.6) was fitted to each of the provinces separately. In the second stage, the correlations defining the matrix ϕ of (2.10) were estimated using a grid search procedure. (See section 2.6). The estimators obtained are, $\text{Diag}(\phi) = (0.5, 0.25, 0.80, 0.0)$. The data used for estimation of the model cover the years 1982-1988. Data for 1989 were used for model diagnostics by comparing the results within and outside the sample period.

3.1 Preliminary Analysis

Table 1 shows a breakdown of the Labour Force in the four provinces by industry. The figures in the table refer to March 1991. The (expected) sample sizes of the LFS are also shown. As can be seen, the percentage breakdowns in the four provinces are very similar justifying the assumption of equal correlations between the error terms of the census models across the provinces. The similarity of the percentage breakdowns suggests also possible improvements in the efficiency of the model based estimators derived from the joint model over estimators which ignore the cross-sectional correlations between the province population values.

Table 1
Labour Force by Industry in the Atlantic Provinces, March 1991

Sample size	Nova Scotia		New Brunswick		Newfoundland		Prince-Edward Island	
	4,409		3,843		2,970		1,421	
	Thousands	%	Thousands	%	Thousands	%	Thousands	%
Agriculture	7	1.7	7	2.3	0.5	0.2	6.0	9.8
Other primary industry	18	4.4	13	4.2	18.0	7.7	4.0	6.6
Manufacturing	44	10.7	37	11.9	23.0	9.9	6.0	9.8
Construction	24	5.9	21	6.8	18.0	7.7	4.0	6.6
Transp. and communication	35	8.6	30	9.6	20.0	8.6	5.0	8.3
Trade and Commerce	81	19.8	61	19.6	41.0	17.6	10.0	16.4
Finance	20	4.9	12	3.9	6.0	2.6	0.5	0.8
Services	143	35.0	107	34.4	83.0	35.6	19.0	31.1
Public Administration	36	8.8	22	7.0	23.0	9.9	6.0	9.8
Unclassified	1	0.2	1	0.3	0.5	0.2	0.5	0.8
Total	409	100.0	311	100.0	233.0	100.0	61.0	100.0

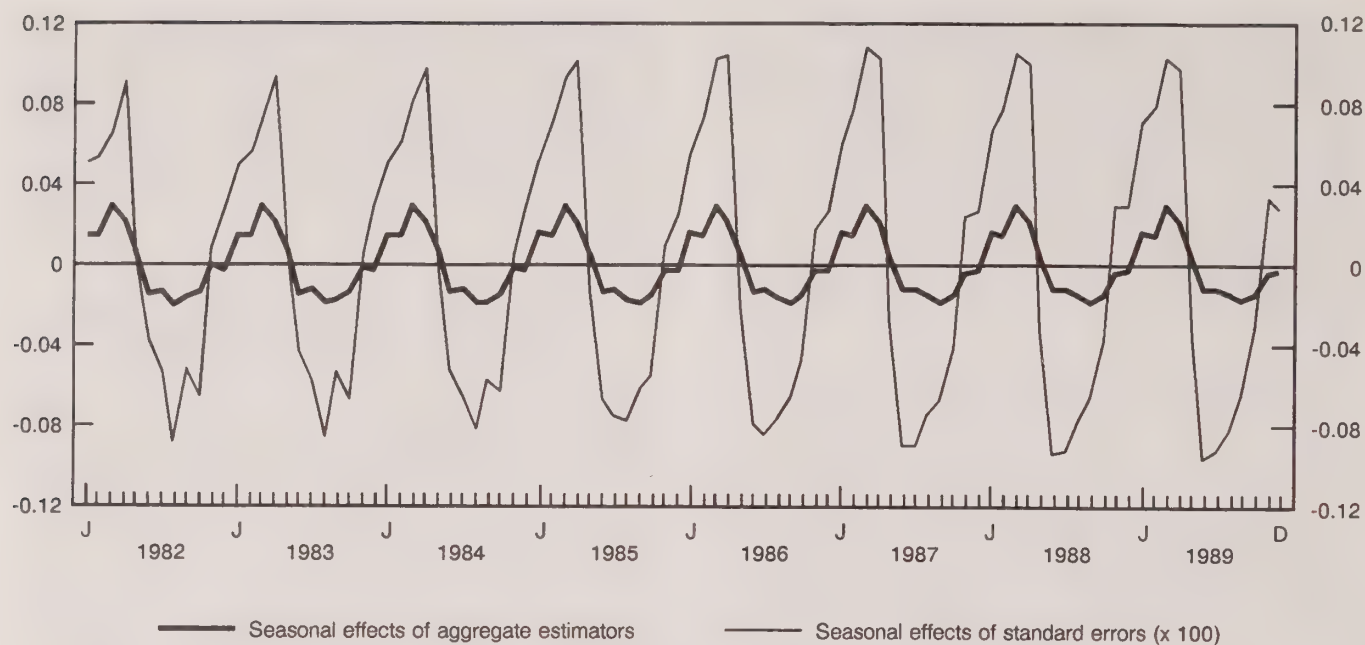


Figure 1. Seasonal Effects of Aggregate Survey Estimators and of Standard Errors of Aggregate Survey Estimators ($\times 100$)

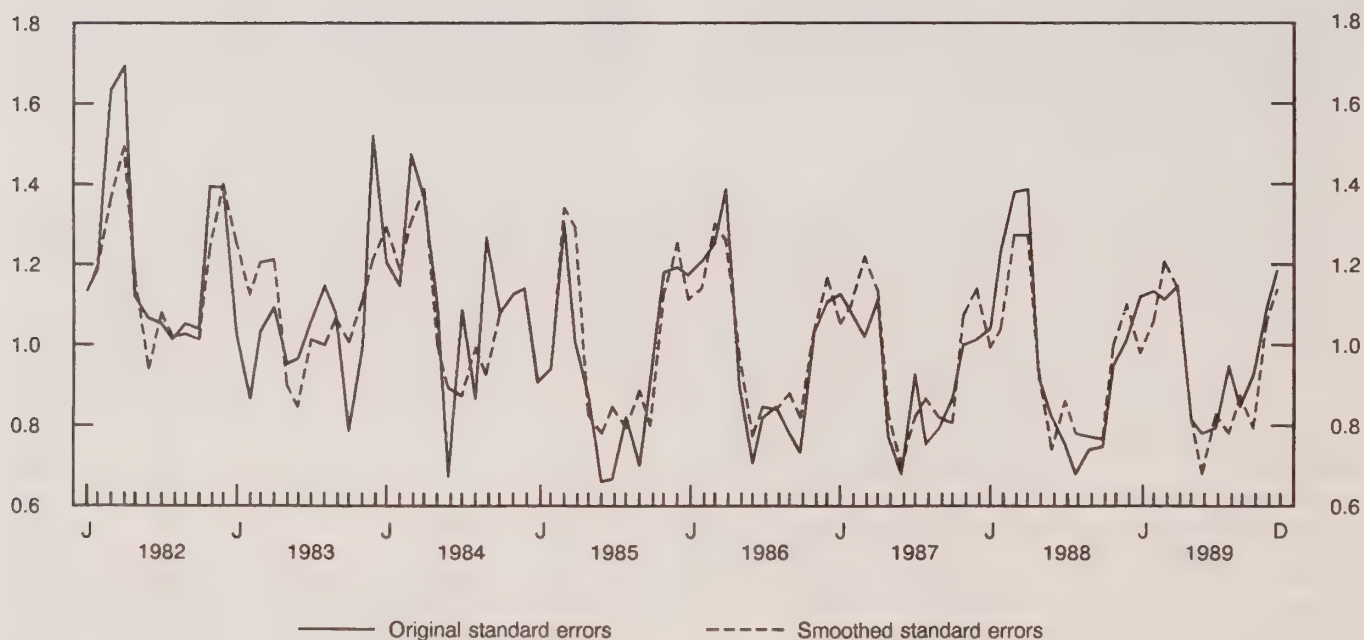


Figure 2. Original and Smoothed Standard Errors of Survey Estimators ($\times 100$) for P.E.I. Province

Two other prior considerations mentioned in section 2.3 are that the model should account for possible rotation group effects and for changes in the variances of the survey errors over time. In order to obtain initial estimates for the rotation group effects, we averaged the pseudo survey errors, $e_{t,p}^{(j)} = (y_t^{(j)} - \bar{y}_t)$, $j = 1, \dots, 6$ over all the months in the sample period. We then divided the averages by the conventional estimates of the standard errors. (The errors $e_{t,p}^{(j)}$ are correlated over time but the correlations are small because except for lags 6, 12 *etc.* the data of any given panel refer to different PSU's in the urban areas and different enumeration areas in the rural areas. See section 2.1). Notice that in the absence of rotation group effects, $E(e_{t,p}^{(j)}) = 0$ for all j and t irrespective of the model postulated for the population values.

This preliminary (model free) analysis yields similar results to the results obtained under the full model, presented in Table 2 of section 3.3.

Next consider the variances of the survey errors.

Figure 1 plots the seasonal effects of the aggregate survey estimators in the four provinces along with the seasonal effects of the standard errors of these estimators (multiplied by 100). Denote as before by w_{ta} the relative labour force size in province a at time t . The aggregate survey estimator is defined as $y_t^* = \sum_{a=1}^4 w_{ta} \bar{y}_{ta}$ (Equation 2.11). The standard error of y_t^* is $(SD^*)_t = [\sum_{a=1}^4 w_{ta}^2 (\widehat{SD}_{ta})^2]^{1/2}$. The seasonal effects were estimated by application of the additive model of X-11 so as not to bind them to any particular model. We chose the additive model since we assume the additive decomposition for the survey estimators. (As revealed from Figure 4, the seasonal effects of the aggregate survey estimators produced by X-11 are very close to the seasonal effects obtained under the model).

Figure 1 shows that the standard errors are influenced by seasonal variations with a seasonal pattern that follows closely the seasonal pattern of the survey estimators and hence of the corresponding population values.

As discussed in section 2.3, rather than using the original estimates of the design standard errors in the models fitted to the panel survey errors we use smoothed values, thus reducing the effect of the sampling errors on the former estimators. Figure 2 plots the two sets of estimators for Prince Edward Island (P.E.I.) province which is the smallest province in the Atlantic region and hence has the smallest sample sizes. As can be seen, the effect of the smoothing is to trim the extreme raw estimates but otherwise the smoothed values behave similarly to the raw estimates. The plots for the other provinces show a similar pattern but the differences between the raw and the smoothed estimates are smaller because of the larger sample sizes in these provinces.

We conclude this section by specifying the models postulated for the seasonal effects in the four provinces. Our initial model assumed fixed variances for the error terms $\eta_{st} = \sum_{j=0}^{11} S_{t+j}$, $t = 1, 2, \dots$ (see equation 2.1). The predicted errors $\hat{\eta}_{st} = \sum_{j=0}^{11} \hat{S}_{t+j}$ obtained under that model were found to decrease in absolute value as a function of time in three out of the four provinces and increase in time in the remaining province. Notice that under the model defined by (2.1), with constant variances of the state error terms, the Kalman filter converges to a steady state by which the V-C matrices of the state vector estimators and hence of $\hat{\eta}_{st}$ are constant. Thus, we modified the initial model such that $\text{VAR}(\eta_{st}) = \sigma_s^2 \times g(t)$ where for the provinces of Nova Scotia, Newfoundland and P.E.I. $g(t) = t^{(-3/2)}$ whereas for New Brunswick $g(t) = t^{1/2}$.

3.2 Results

3.2.1 Rotation Group Biases

Table 2 shows the rotation group Biases (RGB) and their estimated standard errors (SE) in the four provinces as obtained under the full model defined by (2.3), (2.5), (2.6) and (2.10).

Table 2
Rotation Group Biases and Standard Errors
in the Four Provinces ($\times 100$)

Panels	Nova Scotia		New Brunswick		Newfoundland		Prince Edward Island	
	RGB	SE	RGB	SE	RGB	SE	RGB	SE
1	-0.20	0.10	-0.02	0.11	-0.47	0.13	0.32	0.17
2	0.18	0.09	0.40	0.10	0.42	0.12	0.18	0.15
3	0.32	0.08	0.24	0.09	0.47	0.12	0.31	0.15
4	0.06	0.07	0.01	0.09	0.18	0.12	0.03	0.15
5	0.03	0.08	-0.15	0.10	-0.10	0.13	-0.25	0.16
6	-0.34	0.08	-0.50	0.11	-0.50	0.14	-0.60	0.16

The RGB behave fairly consistently across the provinces. Thus, the biases for the 3rd and 6th panel are all highly significant using the conventional t -statistic, having a positive sign for the 3rd panel and a negative sign for the 6th panel. The biases for the 4th and 5th panels have again the same sign in all the provinces and they are all non-significant.

For the 2nd panel all the biases are positive but the bias in P.E.I. is not significant. (P.E.I. is the province with the smallest sample size). It is also in P.E.I. that the sign of the bias for the 1st panel is different from the signs in the other provinces.

As discussed in section 2.3, there is more than one possible reason for the existence of RGB but the results emerging from the Table provide a strong indication that whatever the reason is, the biases found for some of the panels are real and not just the outcome of sampling errors. A drawback of the present analysis, however, is that the RGB are assumed to be fixed over time. Section 4 proposes a more flexible model.

3.2.2 Goodness of Fit

A. TESTING FOR NORMALITY

Let $I_{ta}^{(j)} = (y_{ta}^{(j)} - y_{ta|t-1}^{(j)})$ define the innovation when predicting the j -th panel estimator one month ahead and denote $I'_{ta} = (I_{ta}^{(1)}, \dots, I_{ta}^{(6)})$. The use of maximum likelihood estimation in this study assumes that the vectors I_{ta} are normal deviates (see section 2.4). To test this assumption, we computed the empirical distribution of the standardized innovations $\{(SI)_{ta}^{(j)} = [I_{ta}^{(j)} / \widehat{SD}(I_{ta}^{(j)})], t = (k + 1), \dots, N\}$ and compared it to the standard normal distribution using the Kolmogorov-Smirnov test statistic. This test statistic was computed for each of the six panels in the four provinces yielding P -values larger than 0.15 in 21 out of the 24 cases. (The tests were performed using PROC UNIVARIATE of the SAS package. By this procedure, if the sample size is greater than fifty as it is in our case, the data are tested against a normal distribution with mean and variance equal to the sample mean and variance). Applying the same test procedure to the standardized innovations $\{(SI)_{ta} = [I_{ta} / \widehat{SD}(I_{ta})], t = (k + 1), \dots, N\}$ where $I_{ta} = [\sum_{j=1}^6 I_{ta}^{(j)} / 6]$ yields P -values larger than 0.15 in all the four provinces.

The estimators of the standard deviations of the innovations used for the tests are those produced by the Kalman filter, without accounting for the variance component resulting from parameter estimation (see section 2.5). The

latter component is negligible even in P.E.I. which has the smallest samples sizes among the four provinces. We come back to this finding in section 3.4.

B. PREDICTION ERRORS WITH DIFFERENT PREDICTORS

Table 3 contains summary statistics comparing the behaviour of the prediction errors (innovations) in the four provinces as obtained for three different sets of estimators of the state vectors: (1) The estimators obtained under the separate models (SM) defined by (2.2), (2.5) and 2.6; (2) the estimators obtained under the joint model (JM) defined by (2.2), (2.5), (2.6) and (2.10); (3) the estimators obtained by imposing the robustness constraints (2.11) on the joint model (ROB). Below we define the summary statistics using as before the notation $I_{ta}^{(j)} = (y_{ta}^{(j)} - \hat{y}_{ta|t-1}^{(j)})$ for the prediction error when predicting the j -th panel estimator one month ahead.

$$MB_a = \sum_{t=k+1}^N (\sum_{j=1}^6 I_{ta}^{(j)} / 6) / (N - k) - \text{mean bias in predicting the mean survey estimator} \\ \bar{y}_{ta} = \sum_{j=1}^6 y_{ta}^{(j)} / 6.$$

$$MAB_a = \sum_{j=1}^6 | \sum_{t=k+1}^N I_{ta}^{(j)} / (N - k) | / 6 - \text{mean absolute bias in predicting the panel estimators.}$$

$$SQRE_a = \{ \sum_{t=k+1}^N [1/6 \sum_{j=1}^6 I_{ta}^{(j)} / \bar{y}_{ta}]^2 / (N - k) \}^{1/2} - \text{square root of mean square relative prediction error in predicting the mean survey estimator.}$$

The above summary statistics are shown separately for the sample period of July 1983 – December 1988 and for the postsample period of January 1989 – December 1989. In the latter case, the data were added one data point at a time so that for predicting the survey estimator of February 1989 for example we used the data observed until January 1989 and so forth.

Table 3
Prediction Errors in the Four Provinces,
Summary Statistics ($\times 100$)

	Nova Scotia			New Brunswick			Newfoundland			Prince Edward Island		
	SM	JM	ROB	SM	JM	ROB	SM	JM	ROB	SM	JM	ROB
7.83 – 12.88												
<i>MB</i>	-.11	-.07	-.06	-.12	-.09	-.06	-.25	-.18	-.08	.06	.14	.15
<i>MAB</i>	.12	.11	.10	.14	.12	.11	.29	.24	.20	.20	.23	.23
<i>SQRE</i>	5.76	5.62	5.70	5.48	5.47	5.47	7.03	6.91	6.96	9.34	9.13	9.17
1.89 – 12.89												
<i>MB</i>	.14	.11	.04	.47	.47	.46	.36	.33	.17	.84	.85	.86
<i>MAB</i>	.32	.32	.30	.51	.51	.50	.39	.37	.29	.84	.85	.86
<i>SQRE</i>	6.39	6.27	6.82	6.25	6.25	6.32	5.92	5.90	5.61	9.45	9.26	9.30

The main conclusions from Table 3 are as follows:

- (1) The results obtained for the three sets of predictors are in general very similar, indicating that for the data analyzed the use of the joint model improves only slightly over the use of the separate models and that there are no abrupt changes in the level of the series in the years considered.
- (2) The errors when predicting the survey estimators are small both within and outside the sample period, suggesting a good fit of the model. Notice that except in P.E.I., the relative prediction errors as measured by the statistics $SQRE_a$ are all less than 7%.
- (3) The biases of the prediction errors in the postsample period are larger than in the sample period with relatively large differences in New Brunswick and P.E.I. This outcome by itself could suggest some model failure in the year 1989. Inspection of the monthly panel prediction errors in the four provinces for this year, (not shown in the Table), indicates however that although the errors are in general mostly positive, the relatively large biases are mainly the result of one or two extreme errors which, with only 12 data points, has a large effect on the average summary statistics. It should be noted also that the estimated unemployment

rates in the four provinces in the year 1989 are between 0.11 and 0.18 so that a prediction bias of .005 or even .009 as obtained for P.E.I. is not high. Clearly, the model can be modified to account for these biases if they persist with additional data. On the other hand, notice that the discussion above refers only to the bias of the prediction errors since the bias of the model based estimators of the concurrent population values is controlled by the robustness constraints (2.11).

In view of the very similar results obtained for the three sets of predictors considered and in order to highlight the performance of the robustness constraints, we deliberately deflated the unemployment rates in the period March 1985 to March 1987 by 33%, deflated the rates in the period April 1987 – November 1988 by 25% and inflated the rates in the period December 1988 – December 1989 by 33%. The effect of these operations is to introduce sudden drifts in the data in the months $t = 39$, $t = 64$ and $t = 84$. Figure 3 displays the aggregate, one step ahead prediction errors (APE), $I_t^a = \sum_{a=1}^4 w_{ta} [\sum_{j=1}^6 (y_{ta}^{(j)} - \hat{y}_{ta|t-1}^{(j)}) / 6]$ as obtained for the joint model with and without the robustness constraints, and for the separate models.

The clear conclusion from Figure 3 is that by imposing the constraints, the APE in the periods following the three months with sudden drifts are smaller than the APE

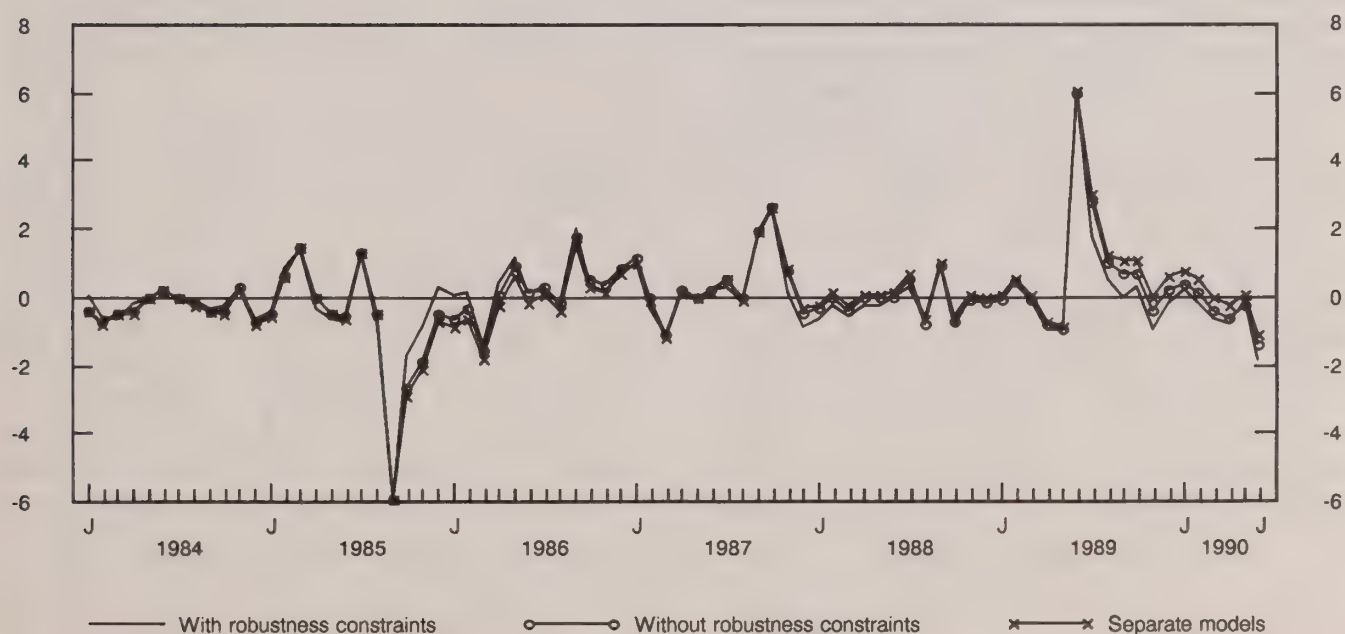


Figure 3. Aggregate One-Step Ahead Prediction Errors of the Three Sets of Predictors ($\times 100$) for Contaminated Data

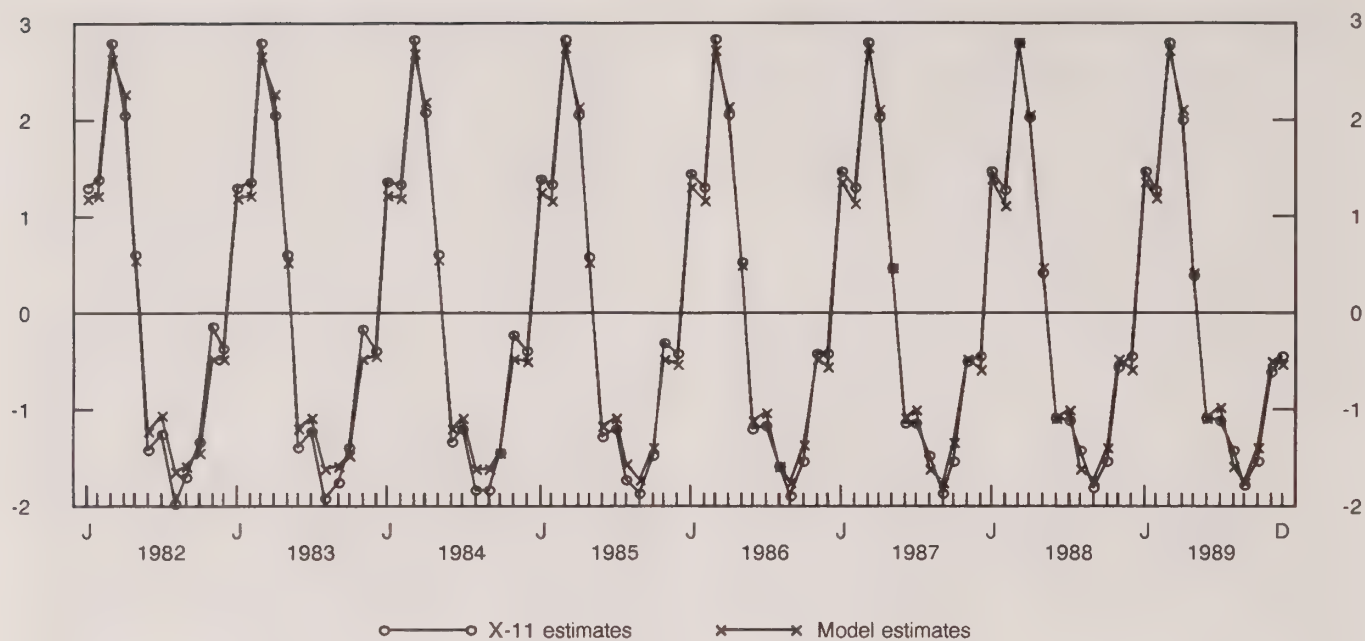


Figure 4. Weighted Averages of Seasonal Effects as Obtained by X-11 and Under the Model ($\times 100$)

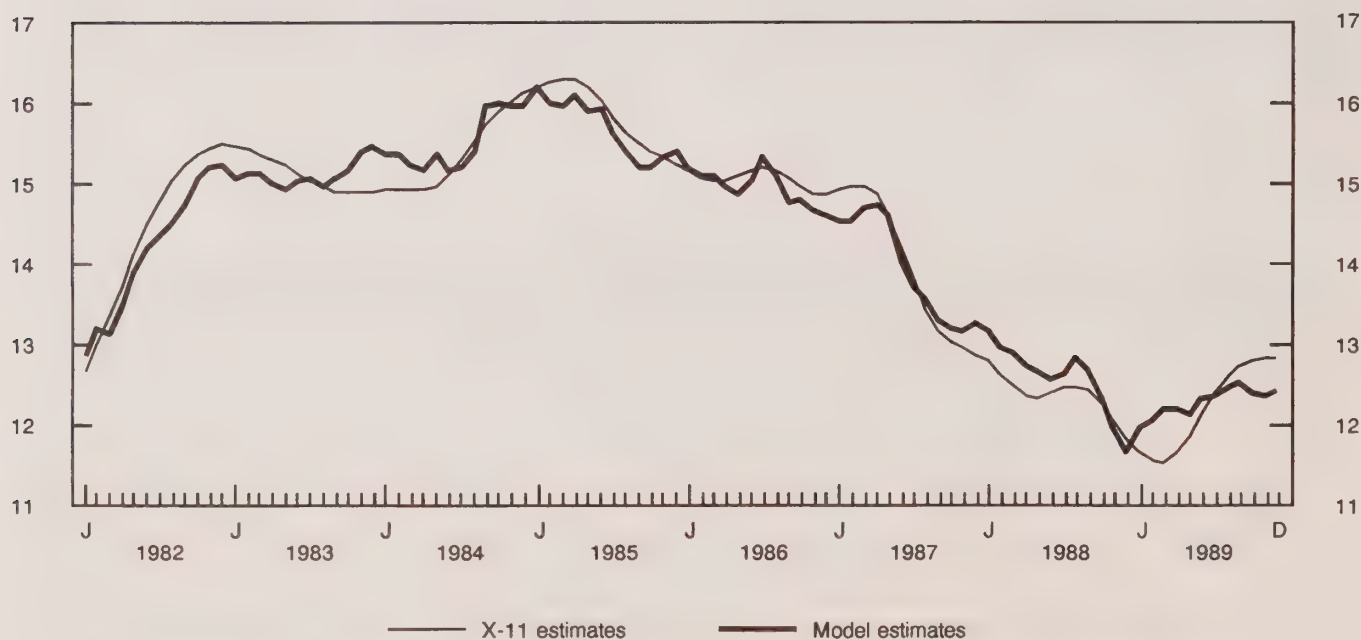


Figure 5. Weighted Averages of Trend Levels as Obtained by X-11 and Under the Model

obtained without the constraints. Thus, in March 1985 for example, $t = 39$, the APE are very large in absolute value both with and without the constraints which is obvious since the predictors use only the data until February 1985. The APE corresponding to the robust predictors return however, to their normal level much faster than the APE of the nonrobust predictors. A similar behaviour is seen to hold in the other two periods. Another notable result featured in the graph is that in the periods following the months with the sudden drifts, the joint model performs better than the separate models even without imposing the robustness constraints. Thus, by borrowing information from one province to the other, the joint model adapts itself more rapidly to the new level of the series. For more illustrations of the performance of the robustness constraints see Pfeiffermann and Burck (1990).

C. COMPARISONS WITH ESTIMATORS PRODUCED BY X-11

As a final assessment of the appropriateness of the model, we compare the estimates of the seasonal effects and the trend levels as obtained under the model, with the estimates produced by the X-11 procedure (Dagum 1980). The latter is known to be less dependent on specific model assumptions. This procedure is the commonly used method for seasonal adjustment throughout the world. Figure 4 displays the average seasonal effects for the four provinces as obtained by X-11 and under the model. Figure 5 displays the corresponding trend level estimates. The averages are computed using the weights (w_{it}) employed in previous analyses. The model based estimates shown in the two figures are the smoothed estimates which, like X-11, employ all the data in the sample period.

As can be seen, the seasonal effects produced by the two approaches are very close. The trend level estimates are also close but the X-11 trend curve is smoother than the model curve. Similar close correspondence between X-11 and the model is obtained for each of the four provinces separately, including, in particular, P.E.I. with its relatively small sample sizes.

3.3 Comparison of Design Based and Model Dependent Estimators

We mention in the introduction that one of the major reasons for wishing to model the raw survey estimators is that the model produces estimates for the population values which, at least in small areas, are more accurate (when the model holds) than the survey estimators. We computed the two sets of estimates for the four provinces and found that as expected, the estimates produced by the two approaches behave very similar but the design based estimators are less stable, having in general higher peaks and lower troughs. An important aspect when comparing

the two sets of estimates is their performance in estimating year to year changes of the population values. Such comparisons are free of the obscuring effects of seasonality. Figure 6 displays the results obtained for P.E.I.. The model dependent estimates are the smoothed values of the joint model which use all the data in all the months. As can be seen, the estimates produced by the model are much more stable and vary only mildly from one month to the other compared to the design based estimates. Figure 7 displays the standard errors (S.E.) of the unemployment rates estimators in P.E.I. as computed under the design, (smoothed values, see Figure 2), and under the joint model. Also shown are the S.E. when fitting the separate model defined by (2.2), (2.5) and (2.6) and the corresponding S.E. after accounting for the use of parameter estimates instead of the unknown parameter values. See section 2.5 for details. (The latter have been computed only for the separate model to save in computing time).

There are three notable features emerging from the graphs:

- (1) The S.E. of the model dependent estimators under the joint model are only mildly smaller than the S.E. obtained for the separate model but considerably smaller than the S.E. of the survey estimators.
- (2) The S.E. of the model dependent estimators behave similarly to the S.E. of the survey estimators, a direct consequence of accounting for the changes in the variances of the survey errors over time in the model. See section 2.3 for details.
- (3) Accounting for the use of estimated parameter values in the computation of the S.E. of the model dependent estimators has only a marginal effect on the computed S.E. Recall that P.E.I. is the province with the smallest sample sizes. The effect of accounting for the use of parameter estimates in the other provinces is even smaller.

4. SUMMARY

This article illustrates that data collected by a complex sampling design, consisting of several stages of selection with rotating panels, can be successfully modelled by a relatively simple model. The model consists of two parts: the census model holding for the population values and the survey errors model describing the time series relationship between the survey errors. The use of the model yields more accurate estimators for the population values and their components like trend and seasonality and it permits estimating the S.E. of these estimators in a rather simple way. The model equations can be modified to secure the robustness of the model-dependent estimators against possible model failures.

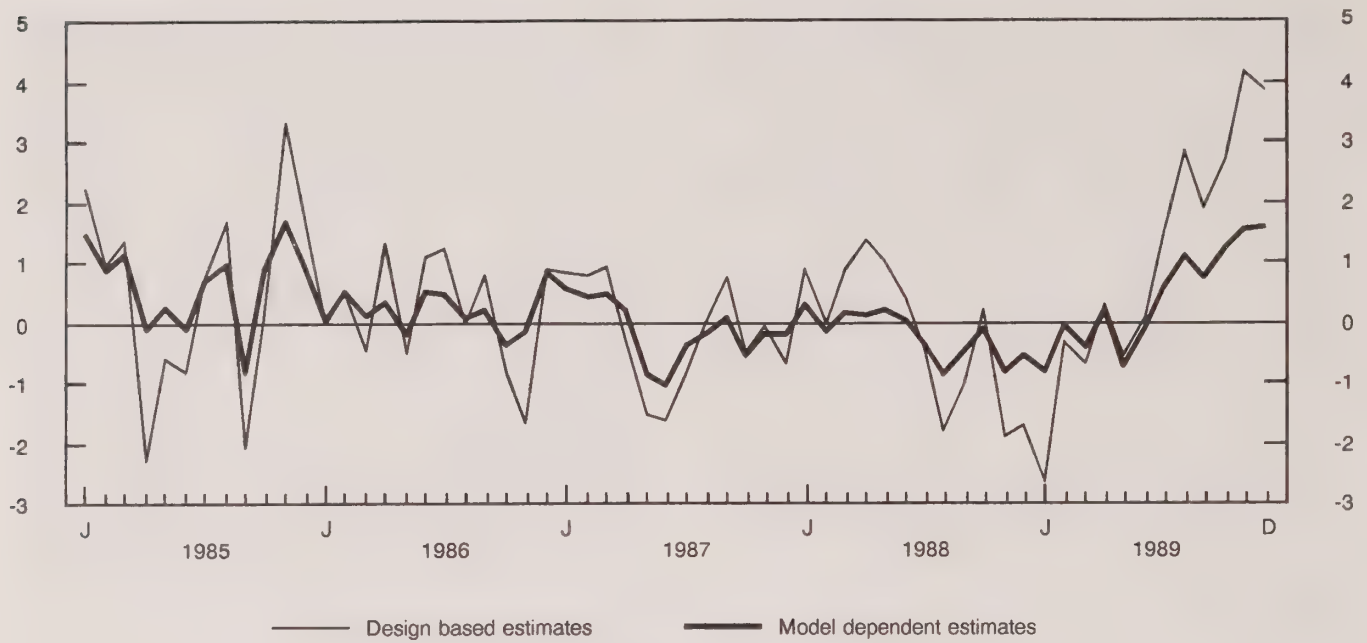


Figure 6. Year to Year Changes in Design Based and Model Dependent Estimates of P.E.I. Unemployment Rates ($\times 100$)

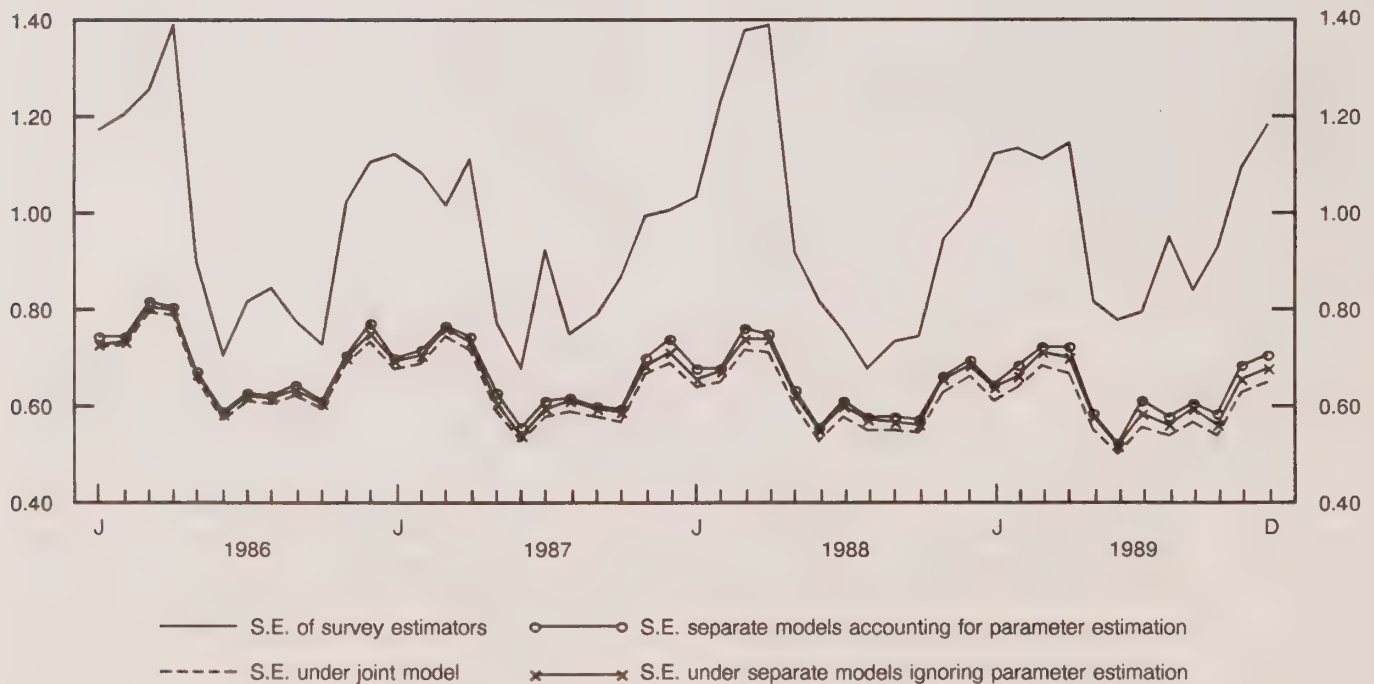


Figure 7. S.E. of Survey Estimators and of Model Dependent Estimators With and Without Accounting for Parameter Estimation ($\times 100$) for P.E.I. Province

The model used in this article can be extended in various directions. Foremost, the model should be applied simultaneously to more provinces or other small areas to ensure that the aggregate sample estimators $\sum_{a=1}^A w_{ta} \bar{y}_{ta}$ are sufficiently close to the corresponding population values. See the discussion in section 2.7. Incorporating in the model an outlier detection mechanism to further assess the performance and suitability of the model is another valuable addition.

Two other extensions are to relax the assumption of constant variance for the error term ϵ_t in the census model and to let the rotation group biases to change over time. The first extension is suggested by the observation made in section 3.1 that the variances of the survey errors are subject to seasonal effects, with a seasonal pattern that is similar to the seasonal pattern of the raw estimates. Fitting the equations (2.4) in the four provinces indicates also the existence of a mild trend in the variances which again behaves similar to the trend of the raw survey estimates. Thus, the variances of the survey errors seem to depend on the magnitude of the survey estimators which suggests that the variances $\sigma_t^2 = V(\epsilon_t)$ change with the level of the population values. As a first approximation one could assume that σ_t^2 is proportional to the corresponding variance of the survey error.

Letting the rotation group biases change over time is a natural extension of the model, considering that the population values means are time dependent. Modelling the evolution of the group biases can however be problematic because of possible identifiability problems with the models holding for the trend and the seasonal effects. See the discussion in Pfeiffermann (1991).

The last two extensions are important and should be explored but based on our experience with the unemployment data, we expect that they will affect the model estimators very mildly.

ACKNOWLEDGEMENT

The authors wish to thank the reviewers for their helpful comments and suggestions. Work on this study was carried out while the first author was staying at Statistics Canada under its Research Fellowship Program.

REFERENCES

- ANSLEY, C.F., and KOHN, R. (1986). Predicted mean square error for state-space models with estimated parameters. *Biometrika*, 73, 467-473.
- BATTESE, G.E., HARTER, R.M., and FULLER, W.A. (1988). An error-components model for prediction of county crop areas using survey and satellite data. *Journal of the American Statistical Association*, 83, 28-36.
- CHANG, I., TIAO, G.C., and CHEN, C. (1988). Estimation of time series parameters in the presence of outliers. *Technometrics*, 30, 193-204.
- DAGUM, E.B. (1980). *The X-11 ARIMA Seasonal Adjustment Method*. Catalogue No. 12-564E, Statistics Canada, Ottawa, Ontario K1A 0T6.
- HAMILTON, J.D. (1986). A standard error for the estimated state vector of a state-space model. *Journal of Econometrics*, 387-397.
- HARRISON, P.J., and STEVENS, C.F. (1976). Bayesian forecasting (with discussion). *Journal of the Royal Statistical Society, Series B*, 38, 205-247.
- HARVEY, A.C. (1984). A unified view of statistical forecasting procedures (with discussion). *Journal of Forecasting*, 3, 245-275.
- HARVEY, A.C., and TODD, P.H.J. (1983). Forecasting economic time series with structural and Box-Jenkins models (with discussion). *Journal of Business and Economic Statistics*, 1, 299-315.
- LEE, H. (1990). Estimation of panel correlations for the Canadian Labour Force Survey. *Survey Methodology*, 16, 283-292.
- MARAVALL, A. (1985). On structural time series models and the characterization of components. *Journal of Business and Economic Statistics*, 3, 350-355.
- MORRIS, N.D., and PFEFFERMANN, D. (1984). A Kalman filter approach to the forecasting of monthly time series affected by moving festivals. *Journal of Time Series*, 5, 255-268.
- PFEFFERMANN, D. (1991). Estimation and seasonal adjustment of population means using data from repeated surveys. *Journal of Business and Economic Statistics*, 9, 163-175.
- PFEFFERMANN, D., and BURCK, L. (1990). Robust small area estimation combining time series and cross-sectional data. *Survey Methodology*, 16, 217-237.
- PFEFFERMANN, D., and BARNARD, C.M. (1991). Some new estimators for small-area means with application to the assessment of farmland values. *Journal of Business and Economic Statistics*, 9, 73-84.
- SINGH, M.P., DREW, J.D., GAMBINO, J.G., and MAYDA, F. (1990). *Methodology of the Canadian Labour Force Survey*. Catalogue No. 71-526, Statistics Canada, Ottawa, Ontario, K1A 0T6.
- TILLER, R.B. (1992). Time series modeling of sample survey data from the U.S. current population survey. *Journal of Official Statistics*, 8, 149-166.

Maximum Likelihood Estimation of Constant Multiplicative Bias Benchmarking Model with Application

IJAZ U.H. MIAN and NORMAND LANIEL¹

ABSTRACT

The maximum likelihood estimation of a non-linear benchmarking model, proposed by Laniel and Fyfe (1989; 1990), is considered. This model takes into account the biases and sampling errors associated with the original series. Since the maximum likelihood estimators of the model parameters are not obtainable in closed forms, two iterative procedures to find the maximum likelihood estimates are discussed. The closed form expressions for the asymptotic variances and covariances of the benchmarked series, and of the fitted values are also provided. The methodology is illustrated using published Canadian retail trade data.

KEY WORDS: Autocorrelations; Bias model; Generalized least squares; Sampling errors.

1. INTRODUCTION

Benchmarking methods are very commonly used for improving sub-annual survey estimates with the help of corresponding estimates, called benchmarks, from an annual survey. The improvement generally is in terms of reductions in the biases and variances of the sub-annual estimates. For example, the monthly retail trade estimates might be improved using estimates from annual retail trade surveys. The sub-annual estimates are often biased due to coverage deficiencies in the frame. Undercoverage is caused by delay in the inclusion of new businesses and non-representation of non-employer businesses in the frame. Furthermore, the variances of the sub-annual estimates are often larger than those of the corresponding annual estimates, and the sampling covariances exist between the estimates from different time periods due to overlap of the samples. On the other hand, the annual estimates can be assumed unbiased because, in practice, their frames do not suffer much from coverage deficiencies. Detailed discussions on benchmarking can be found in Laniel and Fyfe (1989; 1990), Cholette (1987; 1988), and others.

Several procedures for benchmarking time series are available in the literature. Based on a quadratic minimization approach, Denton (1971) proposed several procedures to benchmark a single time series. Cholette (1984) proposed a modified version of Denton's order one proportional variant method where he removed the starting condition to avoid transient effects. The assumptions made by authors are very unlikely to be satisfied by most economic time series. More specifically, their models assume that the bias associated with sub-annual estimates follows a random walk and that both the sub-annual and annual data are observed without sampling errors. In general the estimates come from sample surveys and hence they are subject to sampling errors.

Hillmer and Trabelsi (1987) proposed an alternate approach to benchmarking which is based on an ARIMA model (see *e.g.*, Box and Jenkins 1976). Although this approach takes into account the sampling covariances of the sub-annual and annual estimates, the approach does not accommodate biases in the sub-annual estimates. Cholette and Dagum (1989) modified the Hillmer and Trabelsi approach by replacing the ARIMA model by an "intervention" model. This approach allows the modelling of systematic effects in the time series but still possesses the same weaknesses as found in the Hillmer and Trabelsi model (Laniel and Fyfe 1990).

In order to overcome the deficiencies mentioned above, Laniel and Fyfe (1989; 1990) proposed a non-linear benchmarking model on levels. The authors provided a complex algorithm to find the generalized least squares (GLS) estimates (and their asymptotic covariances) of the model parameters. This model takes into account the sampling covariances of the sub-annual and annual estimates, and can be used when the benchmarks come either from censuses or annual overlapping samples. This model also assumes a constant multiplicative (relative) bias associated with the sub-annual level estimates. Other constant multiplicative bias benchmarking models has been proposed by Cholette (1992) and Laniel and Mian (1991). Cholette assumes a model in which both the bias and errors are multiplicative. The author used the GLS theory to find the estimates of the model parameters after making a logarithmic transformation on the model. Laniel and Mian (1991) have provided an algorithm to find the maximum likelihood estimates of a constant multiplicative bias benchmarking model with mixed (a mixture of binding and non-binding) benchmarks. The binding benchmark here is an estimate from a census (*i.e.*, an estimate with zero variance) and the non-binding benchmark on the other hand is an estimate based on a sample. The assumption

¹ Ijaz U.H. Mian and Normand Laniel, Social Survey Methods Division, Statistics Canada, Ottawa, Ontario, K1A 0T6, Canada.

of a constant multiplicative bias will be verified in practice if the rate of frame maintenance activities is relatively stable, that is, when the proportion of frame coverage deficiencies is fairly constant over time. This assumption also implies that the covered and uncovered businesses in the frame possesses the same average period-to-period ratios with respect to the variable of interest. The nature of bias associated with sub-annual estimates may vary from one time series to another. Cholette and Dagum (1991) have proposed a benchmarking method which assumes a constant additive bias associated with the sub-annual estimates.

The purpose of this paper is to consider the maximum likelihood (ML) estimation of the parameters of Laniel and Fyfe's model and the results are based on the report of Mian and Laniel (1991). Their model is described in the next section. Two iterative processes to find the ML estimates of the model parameters are discussed in Section 3. The closed form expressions for the asymptotic covariances of the estimators of model parameters and of the fitted values are provided in Section 4. The published Canadian retail trade data collected by Statistics Canada are used to illustrate the methodology.

2. CONSTANT MULTIPLICATIVE BIAS MODEL (CMBM)

In order to meet the benchmarking requirements of the economic surveys, the following constant multiplicative bias model (CMBM) has been proposed by Laniel and Fyfe (1989; 1990). The model assumes that the biased sub-annual estimates y_t follow the relationship given by

$$y_t = \beta\theta_t + a_t, \quad t = 1, 2, \dots, n \quad (2.1)$$

and the unbiased annual estimates z_T follow the relationship

$$z_T = \sum_{t \in T} \theta_t + b_T, \quad T = 1, 2, \dots, m, \quad (2.2)$$

where the subscripts t and T denotes respectively the sub-annual and annual time periods, θ_t is the unknown fixed sub-annual parameter, β is an unknown constant bias parameter associated with y_t , and a_t and b_t are sampling errors associated respectively with y_t and z_T . The above model is a hybrid type (mixed) model in which bias is multiplicative but errors are additive.

Before proceeding further, let us define the column vectors $y = (y_1, y_2, \dots, y_n)'$, $z = (z_1, z_2, \dots, z_m)'$, $a = (a_1, a_2, \dots, a_n)'$, $b = (b_1, b_2, \dots, b_m)'$, and $\Theta = (\theta_1, \theta_2, \dots, \theta_n)'$. The CMBM model, given by (2.1) and (2.2), can be rewritten as

$$\begin{aligned} w &= X_\beta \Theta + u \\ &= X_\Theta \beta + X_D \Theta + u, \end{aligned} \quad (2.3)$$

where

$$\begin{aligned} X_\beta &= (\beta I_n : D')', \quad X_\Theta = (\Theta' : \mathbf{0}')', \quad X_D = (\mathbf{0}' : D')', \\ w &= (y' : z')', \quad u = (a' : b')', \quad D = (d_{Tt}), \end{aligned} \quad (2.4)$$

I_n is an identity matrix of order n , $\mathbf{0}$ is a zero vector or matrix of an appropriate order, and d_{Tt} is an indicator function equal to 1 for $t \in T$ and to 0 otherwise. It is assumed that the sampling error vectors a and b follow multivariate normal distributions such that $a \sim MN(\mathbf{0}, V_{aa})$ and $b \sim MN(\mathbf{0}, V_{bb})$. Also, in the general case, a and b are correlated, which means that $\text{Cov}(a, b) = V_{ab} = V'_{ba} \neq \mathbf{0}$. It is shown in the next section that the ML and GLS estimators of the Θ and β are same for this model. Thus the assumption regarding the normality of a and b is required only to obtain the Fisher information matrix (and hence variances) of the ML estimators.

3. MAXIMUM LIKELIHOOD ESTIMATION

The log-likelihood function under CMBM can be written as

$$\ln(L) = -\frac{(n+m)}{2} \ln(2\pi) - \frac{1}{2} \ln |V| - \frac{1}{2} Q, \quad (3.1)$$

where

$$Q = (w - X_\beta \Theta)' V^{-1} (w - X_\beta \Theta) \quad (3.2)$$

and

$$V = \begin{pmatrix} V_{aa} & V_{ab} \\ V_{ba} & V_{bb} \end{pmatrix}.$$

The ML estimates of the model parameters Θ and β can be obtained, assuming V known, by maximizing the log-likelihood function (3.1) or equivalently by minimizing the quadratic term Q (3.2). For this particular model, the ML and GLS estimators of the model parameters are the same and the distinction between them will be made only if the need arises. Taking the first order partial derivatives of $\ln(L)$ with respect to Θ and β , respectively, and then equating them to zero, we have

$$\begin{aligned} \frac{\partial \ln(L)}{\partial \Theta} &= X'_\beta V^{-1} (w - X_\beta \Theta) = \mathbf{0}, \\ \frac{\partial \ln(L)}{\partial \beta} &= X'_\Theta V^{-1} (w - X_\beta \Theta) = 0. \end{aligned} \quad (3.3)$$

Since $E(\mathbf{w}) = \mathbf{X}_\beta \boldsymbol{\Theta}$ under the model (2.3), the above equations are estimating equations in the sense of Godambe (1960) and they are information unbiased. It is interesting to note that $\mathbf{X}'_\beta \mathbf{V}^{-1}$ and $\mathbf{X}'_0 \mathbf{V}^{-1}$ do not depend on \mathbf{w} so that the equations (3.3) converge to zeros and hence have consistent roots as long as $E(\mathbf{w}) = \mathbf{X}_\beta \boldsymbol{\Theta}$. That is, even when \mathbf{V} in the above equations is replaced by some of its consistent estimate the equations will provide consistent estimates of the vector $\boldsymbol{\Theta}$ and β . Also note that the above equations are non-linear in the parameters to be estimated and it is not possible to obtain explicit expressions for the estimators of $\boldsymbol{\Theta}$ and β . Therefore some iterative procedure, such as the well-known Fisher-Newton-Raphson method (also called method of scores by Fisher), may be used to obtain the estimates. The elements of expected Fisher information matrix needed to implement the Fisher-Newton-Raphson method are provided in Section 4.

An alternate way to find the ML estimates of the model parameters is to solve the estimating equations (3.3) successively. By solving the first expression of (3.3), the estimate of $\boldsymbol{\Theta}$, as a function of β , is given by

$$\hat{\boldsymbol{\Theta}} \equiv \hat{\boldsymbol{\Theta}}(\beta) = (\mathbf{X}'_\beta \mathbf{V}^{-1} \mathbf{X}_\beta)^{-1} \mathbf{X}'_\beta \mathbf{V}^{-1} \mathbf{w}. \quad (3.4)$$

Similarly, by solving the second expression of (3.3), the estimator of β , as a function of $\boldsymbol{\Theta}$, is given by

$$\hat{\beta} \equiv \hat{\beta}(\boldsymbol{\Theta}) = [\boldsymbol{\Theta}' \mathbf{V}_{aa.b}^{-1} (\mathbf{y} - \mathbf{V}_{ab} \mathbf{V}_{bb}^{-1} (\mathbf{z} - \mathbf{D}\boldsymbol{\Theta}))] / [\boldsymbol{\Theta}' \mathbf{V}_{aa.b}^{-1} \boldsymbol{\Theta}], \quad (3.5)$$

where

$$\mathbf{V}_{aa.b} = \mathbf{V}_{aa} - \mathbf{V}_{ab} \mathbf{V}_{bb}^{-1} \mathbf{V}_{ba}.$$

The ML estimates of $\boldsymbol{\Theta}$ and β can be obtained by successively calculating equations (3.4) and (3.5) until convergence. This procedure has an advantage over the Fisher-Newton-Raphson method as it is easy to implement. However, for this kind of algorithm, the convergence is usually very slow. We will compare these two methods in Section 6 to check the speed of their convergence.

Once the ML estimates of the model parameters are obtained, one can find the fitted sub-annual values $\hat{\mathbf{y}} = \hat{\beta} \hat{\boldsymbol{\Theta}}$ and the fitted annual values $\hat{\mathbf{z}} = \mathbf{D} \hat{\boldsymbol{\Theta}}$.

Initial Guess for $\boldsymbol{\Theta}$ and β

In order to obtain an initial guess for β , say $\hat{\beta}_0$, let us rewrite the model (2.3) as

$$\mathbf{w}^* = \mathbf{X}_\Theta^* \beta + \mathbf{u}^*,$$

where $\mathbf{w}^* = ((\mathbf{D}\mathbf{y})' : (\mathbf{z} - \mathbf{D}\boldsymbol{\Theta})')'$, $\mathbf{X}_\Theta^* = ((\mathbf{D}\boldsymbol{\Theta})' : \mathbf{0}')$ and $\mathbf{u}^* = ((\mathbf{D}\mathbf{a})' : \mathbf{b}')$. Thus the ML estimate of β is given by

$$\hat{\beta} = [\mathbf{X}_\Theta^{*'} (\mathbf{V}^*)^{-1} \mathbf{w}^*] / [\mathbf{X}_\Theta^{*'} (\mathbf{V}^*)^{-1} \mathbf{X}_\Theta^*], \quad (3.6)$$

where

$$\mathbf{V}^* = \text{Cov}(\mathbf{u}^*) = \begin{pmatrix} \mathbf{D}\mathbf{V}_{aa} \mathbf{D}' & \mathbf{D}\mathbf{V}_{ab} \\ \mathbf{V}_{ba} \mathbf{D}' & \mathbf{V}_{bb} \end{pmatrix}.$$

Using the fact that $E(\mathbf{z}) = \mathbf{D}\boldsymbol{\Theta}$, and replacing $\mathbf{D}\boldsymbol{\Theta}$ by \mathbf{z} in (3.6), an initial guess for β may be taken as

$$\begin{aligned} \hat{\beta}_0 &= \left[\begin{pmatrix} \mathbf{z} \\ \mathbf{0} \end{pmatrix}' (\mathbf{V}^*)^{-1} \mathbf{w}^* \right] / \left[\begin{pmatrix} \mathbf{z} \\ \mathbf{0} \end{pmatrix}' (\mathbf{V}^*)^{-1} \begin{pmatrix} \mathbf{z} \\ \mathbf{0} \end{pmatrix} \right] \\ &= [\mathbf{z}' (\mathbf{D}\mathbf{V}_{aa.b} \mathbf{D}')^{-1} \mathbf{D}\mathbf{y}] / [\mathbf{z}' (\mathbf{D}\mathbf{V}_{aa.b} \mathbf{D}')^{-1} \mathbf{z}]. \end{aligned} \quad (3.7)$$

The initial estimate of $\boldsymbol{\Theta}$ can be obtained from (3.4) by replacing β by $\hat{\beta}_0$.

4. COVARIANCES OF THE ESTIMATORS

In this section, we derive the expressions for the asymptotic covariances of the ML estimators of CMBM parameters by inverting the Fisher information matrix, say $\boldsymbol{\Omega}$. The asymptotic covariances of the fitted sub-annual and annual values are provided by using the delta method. First, let us consider the derivation of the covariances of the ML estimators of $\boldsymbol{\Theta}$ and β . The elements of $\boldsymbol{\Omega}$ (i.e., the negative expectations of the second order partial derivatives of $\ln(L)$) are given by

$$\Omega_{11} = -E \left[\frac{\partial^2 \ln(L)}{\partial \boldsymbol{\Theta} \partial \boldsymbol{\Theta}'} \right] = \mathbf{X}'_\beta \mathbf{V}^{-1} \mathbf{X}_\beta,$$

$$\Omega_{22} = -E \left[\frac{\partial^2 \ln(L)}{\partial \beta^2} \right] = \boldsymbol{\Theta}' \mathbf{V}_{aa.b}^{-1} \boldsymbol{\Theta}$$

and

$$\Omega_{12} = \Omega_{21} = -E \left[\frac{\partial^2 \ln(L)}{\partial \boldsymbol{\Theta} \partial \beta} \right] = \mathbf{X}'_\beta \mathbf{V}^{-1} \mathbf{X}_\Theta.$$

Therefore, the Fisher information matrix of order $(n + 1) \times (n + 1)$ is given by

$$\boldsymbol{\Omega} = \begin{bmatrix} \Omega_{11} & \Omega_{12} \\ \Omega_{21} & \Omega_{22} \end{bmatrix}. \quad (4.1)$$

Inverting Ω by using the algebra of partitioned matrices we have

$$\begin{aligned}\text{Cov}(\hat{\Theta}) &= \Omega_{11.2}^{-1}, \\ \text{Var}(\hat{\beta}) &= \Omega_{22.1}^{-1}, \\ \text{Cov}(\hat{\Theta}, \hat{\beta}) &= -\Omega_{11.2}^{-1} \Omega_{12} \Omega_{22}^{-1} \\ &= -\Omega_{11}^{-1} \Omega_{12} \Omega_{22.1}^{-1},\end{aligned}\quad (4.2)$$

where

$$\begin{aligned}\Omega_{11.2} &= \Omega_{11} - \Omega_{12} \Omega_{22}^{-1} \Omega_{21}, \\ \Omega_{22.1} &= \Omega_{22} - \Omega_{21} \Omega_{11}^{-1} \Omega_{12}.\end{aligned}\quad (4.3)$$

Once the covariance matrix Ω^{-1} is available, the asymptotic covariances of the sub-annual fitted values \hat{y} can be obtained by using the delta method (see *e.g.*, Rao 1973). Let Δ be the matrix of first order partial derivatives of y with respect to the elements of $(\Theta':\beta)'$. Clearly, the $n \times (n+1)$ matrix is $\Delta = (\beta I_n : \Theta)$. Now, by using the delta method, the asymptotic covariance matrix of \hat{y} is given by

$$\text{Cov}(\hat{y}) = \Delta \Omega^{-1} \Delta'. \quad (4.4)$$

Furthermore, the covariance matrix of the annual fitted values \hat{z} , from the standard multivariate normal theory, is given by

$$\text{Cov}(\hat{z}) = D \Omega_{11.2}^{-1} D', \quad (4.5)$$

where D and $\Omega_{11.2}$ are as defined by (2.4) and (4.3), respectively.

5. MAXIMUM LIKELIHOOD ESTIMATION WHEN $V_{ab} = 0$

In this section we consider the ML estimation of the model parameters for the special case when the error vectors a and b are uncorrelated (*i.e.*, $\text{Cov}(a, b) = V_{ab} = V'_{ba} = 0$). Usually this is the case in sample surveys when annual and sub-annual samples are drawn independently from each other. Reduction in the results of Sections 3 and 4 can be seen by substituting $V_{ab} = V'_{ba} = 0$ in the equations. As an example, for this special case, the ML estimators of Θ and β , given by (3.4) and (3.5), reduce to

$$\begin{aligned}\hat{\Theta}^* &\equiv \hat{\Theta}^*(\beta) = (\beta^2 V_{aa}^{-1} + D' V_{bb}^{-1} D)^{-1} \\ &\quad (\beta V_{aa}^{-1} y + D' V_{bb}^{-1} z)\end{aligned}$$

and

$$\hat{\beta}^* \equiv \hat{\beta}^*(\Theta) = [\Theta' V_{aa}^{-1} y] / [\Theta' V_{aa}^{-1} \Theta],$$

respectively. These equations must be solved successively to obtain the required estimates.

Similarly, the elements of the Fisher information matrix reduce to

$$\Omega_{11}^* = \beta^2 V_{aa}^{-1} + D' V_{bb}^{-1} D,$$

$$\Omega_{22}^* = \Theta' V_{aa}^{-1} \Theta,$$

$$\Omega_{12}^* = \Omega_{21}^{*'} = \beta V_{aa}^{-1} \Theta.$$

6. AN APPLICATION

Here we present an example using published Canadian retail trade data which results from monthly and annual retail trade surveys conducted by Statistics Canada. The monthly retail trade estimates and their coefficients of variation (CV) are available from the Statistics Canada publication "Retail Trade" (Catalogue No. 63-005 Monthly). There are two types of monthly retail trade estimates, namely preliminary and revised estimates. We use the revised but seasonally unadjusted (raw) estimates for this example. Since the CVs of the revised estimates are not available, the CVs of the preliminary estimates are used to approximate the variances of the revised monthly estimates. The data for the period January 1985 to December 1988 are used in this example. Another difficulty was to find the autocorrelations for monthly retail trade estimates. Based on some monthly retail trade data, Hidioglou and Giroux (1986) provided the estimates of autocorrelations at lags 1, 3, 6, 9 and 12 for three different kinds of stratum in several provinces of Canada. As an approximation to the autocorrelations of monthly retail trade estimates, the averages of their estimates of autocorrelations for the strata in the Province of Ontario and Standard Industrial Classification Code 60 (Foods, Beverages, and Drug industries) are used. The approximate (averaged) autocorrelations, say $\rho(k)$, are given in Table 1.

Table 1
Approximate Autocorrelations $\rho(k)$ for the Monthly
Retail Trade Estimates

Lag k	1	3	6	9	12
$\rho(k)$	0.970	0.940	0.918	0.914	0.962

The method of ordinary least squares and an algorithm of McLeod (1975) for the derivation of theoretical autocorrelations for autoregressive moving-average time series was used to revise the observed autocorrelations. An ARMA (1,0)(1,0)₁₂ seasonal multiplicative model was fitted on the five observed autocorrelations by minimizing the sum of squared differences between the observed and theoretical autocorrelations. Then the estimated model parameters and the above mentioned algorithm of McLeod were used to calculate the autocorrelations for all other lags of interest. Given that the ARMA model is correct for theoretical autocorrelations, this approach provides a consistent estimate of the autocorrelation function. These final (revised) approximate autocorrelations for up to 47 lags are given in Table 2 and were used to approximate the covariances for monthly retail trade estimates via multiplication with the standard deviations.

Table 2

Revised Approximate Autocorrelations $\rho^*(k)$ for the Monthly Retail Trade Estimates for up to 47 Lags

Lag k	$\rho^*(k)$	Lag k	$\rho^*(k)$	Lag k	$\rho^*(k)$	Lag k	$\rho^*(k)$
0	1.0000	12	0.9602	24	0.8896	36	0.8100
1	0.9758	13	0.9345	25	0.8647	37	0.7869
2	0.9555	14	0.9126	26	0.8433	38	0.7669
3	0.9391	15	0.8943	27	0.8253	39	0.7501
4	0.9266	16	0.8798	28	0.8107	40	0.7363
5	0.9177	17	0.8687	29	0.7994	41	0.7254
6	0.9126	18	0.8612	30	0.7913	42	0.7176
7	0.9113	19	0.8572	31	0.7864	43	0.7126
8	0.9136	20	0.8567	32	0.7843	44	0.7106
9	0.9196	21	0.8595	33	0.7862	45	0.7114
10	0.9293	22	0.8661	34	0.7909	46	0.7151
11	0.9429	23	0.8760	35	0.7989	47	0.7217

At the time this study was performed, the annual retail trade estimates were only available for years 1985 through 1988. These estimates are available from Statistics Canada publication "Annual Retail Trade" (Catalogue No. 63-223 Annual). The variances of annual retail trade estimates are not available from the literature and have been computed from the actual survey data. The covariances between monthly and annual estimates are zero because the samples of monthly and annual retail trade surveys were drawn independently from each other. The annual retail trade estimates are from dependent samples, thus their covariances are non-zero. But the estimates of covariances are not readily available via regular survey processing and

a study would be required to obtain them. Consequently, for the purpose of this example, we assumed that the covariances between annual retail trade estimates are zero.

An interesting question was raised by one of the referees. He asked what will happen when the variances and covariances of survey estimates are not known. This is a difficult problem and cannot be answered so easily. However the model presented assumes these variances and covariances are known. In general, the estimating equations used to find the maximum likelihood estimates need only the consistent estimates of variances and covariances. It is a common practice in benchmarking problems to estimate these variances and covariances from survey data since the theoretical values are never known (see, e.g., Hillmer and Trabelsi 1987).

The computations required for this example are performed by an algorithm written in the GAUSS programming language for micro computers. The initial estimate of β for the iterative process, obtained from (3.7), is given by $\hat{\beta}_0 = 0.9162$. The initial estimate of the parameter vector Θ is obtained from (3.4), after replacing β by $\hat{\beta}_0$. Both the Fisher-Newton-Raphson and successive iteration methods, as discussed in Section 3, are used to find the ML estimates of the model parameters. The final ML estimate of β is found to be very close to the initial estimate and is given by $\hat{\beta} = 0.9016$ with $CV = 0.0065$. It is interesting to note that the Fisher-Newton-Raphson method converged very quickly to a final solution for this example. In fact it converged in only 6 iterations (about 1 minute) for a ten digit precision whereas the successive calculations method converged, with the same precision, in over 500 iterations (over 45 minutes), on a 386DX-25Mhz personal computer. However, as they should, both methods converged to the same final solution. The covariance matrix of the estimated vector $(\hat{\Theta}':\hat{\beta})'$ is obtained by inverting the Fisher information matrix Ω , given by (4.1), after replacing parameters by their ML estimates. The original series of the monthly retail trade estimates and the benchmarked series of the ML estimates along with their CVs are given in Table 3. The fitted sub-annual series along with their CVs are also given in this table (last two columns). The original and benchmarked series are also plotted in Figure 1. The results show that the original behaviour of the series is not disturbed by benchmarking and a very large reduction in the CVs of sub-annual estimates is achieved. The original series of the annual retail trade estimates and fitted annual values along with their CVs are given in Table 4. The variances of the fitted values in Tables 3 and 4 are obtained by using expressions (4.4) and (4.5), respectively, after replacing parameters by their ML estimates. The results of fitted values also show a large reduction in the CV's of the original estimates. That is, the reliability of the monthly and annual series are increased by benchmarking.

Table 3
 Monthly Retail Trade Estimates, ML Estimates of the Θ_t 's and Fitted Values
 (all in millions of dollars) Along with their CV's

Year	Month	y_t^*	$CV(y_t)^*$	$\hat{\Theta}_t$	$CV(\hat{\Theta}_t)$	\hat{y}_t	$CV(\hat{y}_t)$
1985	1	8,689.668	0.008	9,686.630	0.00210	8,733.384	0.00667
	2	8,390.380	0.008	9,350.078	0.00210	8,429.951	0.00665
	3	10,107.485	0.006	11,248.048	0.00233	10,141.146	0.00496
	4	10,541.145	0.008	11,741.785	0.00200	10,586.294	0.00656
	5	11,763.659	0.007	13,094.151	0.00198	11,805.576	0.00570
	6	11,067.487	0.008	12,321.326	0.00189	11,108.803	0.00647
	7	10,810.755	0.008	12,029.467	0.00184	10,845.666	0.00643
	8	11,289.656	0.009	12,554.808	0.00206	11,319.309	0.00726
	9	10,336.540	0.009	11,484.216	0.00205	10,354.073	0.00728
	10	11,213.751	0.010	12,447.696	0.00256	11,222.737	0.00809
	11	11,935.495	0.010	13,234.412	0.00258	11,932.034	0.00808
	12	13,300.288	0.008	14,734.891	0.00188	13,284.853	0.00643
1986	1	9,753.373	0.009	10,794.009	0.00221	9,731.787	0.00716
	2	9,249.279	0.009	10,227.777	0.00224	9,221.277	0.00709
	3	10,609.952	0.008	11,729.293	0.00207	10,575.031	0.00622
	4	11,637.936	0.008	12,860.626	0.00206	11,595.032	0.00614
	5	12,695.108	0.008	14,024.139	0.00205	12,644.046	0.00605
	6	11,826.254	0.008	13,059.556	0.00202	11,774.385	0.00598
	7	11,940.908	0.010	13,164.500	0.00233	11,869.002	0.00740
	8	11,866.547	0.010	13,070.205	0.00232	11,783.987	0.00743
	9	11,540.397	0.009	12,712.283	0.00202	11,461.287	0.00670
	10	12,208.845	0.010	13,430.932	0.00235	12,109.215	0.00747
	11	12,201.498	0.010	13,418.219	0.00240	12,097.753	0.00747
	12	14,479.170	0.009	15,933.951	0.00215	14,365.916	0.00670
1987	1	10,271.723	0.012	11,276.676	0.00357	10,166.956	0.00891
	2	9,951.105	0.010	10,945.319	0.00261	9,868.208	0.00737
	3	11,492.162	0.008	12,663.849	0.00230	11,417.620	0.00584
	4	12,867.443	0.009	14,172.605	0.00235	12,777.901	0.00652
	5	13,508.434	0.012	14,850.145	0.00343	13,388.765	0.00862
	6	13,608.274	0.011	14,973.985	0.00287	13,500.418	0.00786
	7	13,278.474	0.023	14,483.340	0.01066	13,058.057	0.00165
	8	12,728.196	0.008	14,028.998	0.00227	12,648.426	0.00577
	9	12,616.239	0.009	13,888.982	0.00233	12,522.188	0.00659
	10	13,760.829	0.008	15,156.409	0.00227	13,664.890	0.00592
	11	13,380.142	0.008	14,733.240	0.00227	13,283.365	0.00597
	12	16,269.757	0.007	17,928.148	0.00241	16,163.867	0.00525
1988	1	11,134.013	0.010	12,234.529	0.00274	11,030.548	0.00753
	2	10,959.374	0.010	12,042.761	0.00276	10,857.651	0.00754
	3	13,177.788	0.008	14,508.565	0.00233	13,080.800	0.00602
	4	13,666.311	0.009	15,035.737	0.00243	13,556.094	0.00676
	5	14,267.530	0.006	15,742.039	0.00379	14,192.890	0.00448
	6	14,432.944	0.009	15,884.130	0.00240	14,320.997	0.00673
	7	13,960.825	0.009	15,363.957	0.00240	13,852.014	0.00673
	8	13,691.315	0.008	15,073.691	0.00233	13,590.312	0.00606
	9	13,773.109	0.008	15,159.075	0.00235	13,667.294	0.00613
	10	13,900.743	0.009	15,279.950	0.00255	13,776.282	0.00696
	11	14,453.461	0.009	15,884.279	0.00260	14,321.132	0.00700
	12	17,772.990	0.009	19,529.791	0.00267	17,607.895	0.00702

*Source: Statistics Canada publication "Retail Trade" (Catalogue No. 63-005 Monthly).

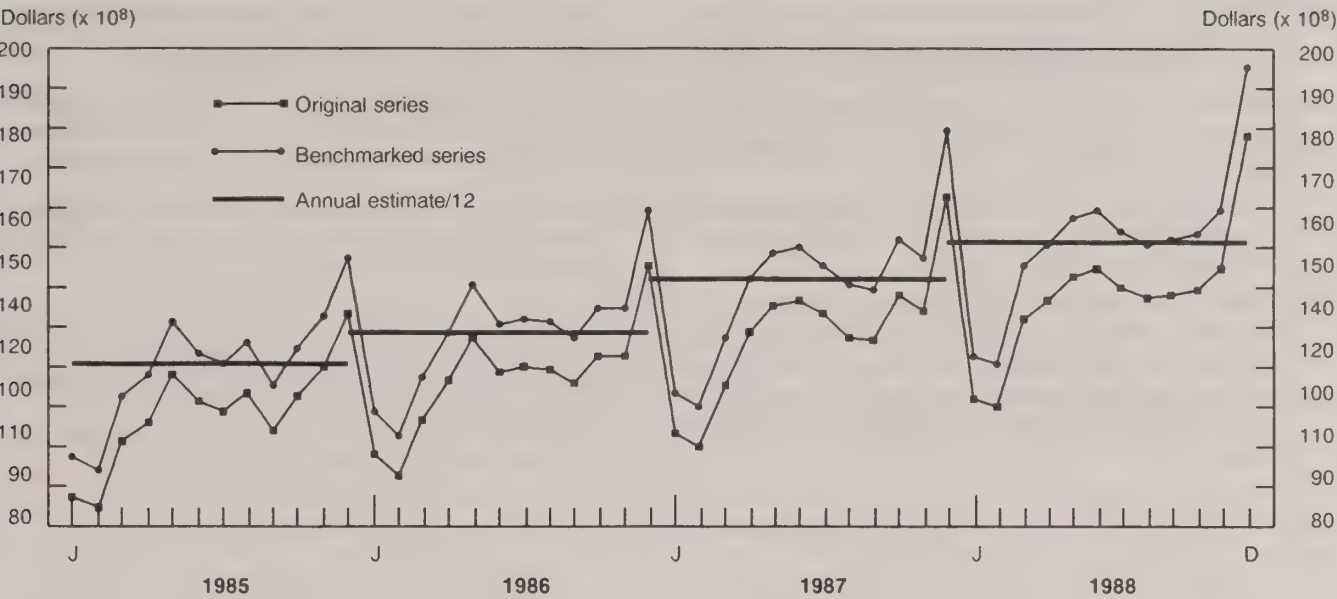


Figure 1. Original and Benchmarked Series of Monthly Retail Trade Estimates for All Stores in Canada

Table 4

Annual Retail Trade Estimates and Annual Fitted Values (in millions of dollars) Along with their CV's

Year	z_T^*	$CV(z_T)$	\hat{z}_T	$CV(\hat{z}_T)$
1985	143,965.400	0.00033	143,927.507	0.00032
1986	154,377.100	0.00031	154,425.491	0.00030
1987	169,944.600	0.00193	169,101.697	0.00128
1988	181,594.000	0.00137	181,738.512	0.00127

*Source: Statistics Canada publication "Annual Retail Trade" (Catalogue No. 63-223 Annual).

7. CONCLUSIONS

The non-linear model discussed here seems to be very appropriate for benchmarking an economic time series from large sample surveys. The proposed iterative procedures to find the maximum likelihood estimates of the model parameters are very simple to implement in practice. However, the convergence of the successive calculation method is very slow in comparison to the Fisher-Newton-Raphson method. The closed form expressions for the covariances of the ML estimators are provided. These estimates and their covariances may be used to make inferences regarding model parameters. Furthermore, expressions for the fitted sub-annual and annual values along with their asymptotic covariances are also provided. The methodology presented in this article seems to provide a good fit to the Canadian retail trade data. However, the goodness of fit tests for this and other benchmarking models need to be developed.

ACKNOWLEDGEMENTS

We acknowledge the valuable suggestions made by the editor, associate editor and two referees which had greatly improved the quality of this paper. We are also thankful to Drs. J. Gambino and M. Kovačević of Statistics Canada for their comments on this research.

REFERENCES

BOX, G.E.P., and JENKINS, G.M. (1976). *Time Series Analysis, Forecasting and Control*. New York: Holden-Day.

CHOLETTE, P.A. (1984). Adjusting sub-annual series to yearly benchmarks. *Survey Methodology*, 10, 35-49.

CHOLETTE, P.A. (1987). Benchmarking and interpolation of time series. Methodology Branch Working Paper, Statistics Canada.

CHOLETTE, P.A. (1988). Benchmarking systems of socio-economic time series. Methodology Branch Working Paper, Statistics Canada.

CHOLETTE, P.A. (1992). Users' manual of programmes BENCH and CALEND to benchmark, interpolate and Calendarize time series data on micro computers. Methodology Branch Working Paper, Statistics Canada.

CHOLETTE, P.A., and DAGUM, E.B. (1989). Benchmarking socio-economic time series data: a unified approach. Methodology Branch Working Paper, Statistics Canada.

CHOLETTE, P.A., and DAGUM, E.B. (1991). Benchmarking time series with autocorrelated sampling errors. Methodology Branch Working Paper, Statistics Canada.

- DENTON, F.T. (1971). Adjustment on monthly or quarterly series to annual totals: An approach based on quadratic minimization, *Journal of the American Statistical Association*, 66, 99-102.
- GODAMBE, V.P. (1960). An optimum property of regular maximum likelihood estimation. *Annals of Mathematical Statistics*, 13, 1208-1211.
- HIDIROGLOU, M.A., and GIROUX, S. (1986). Composite estimation for the Retail Trade Survey. Methodology Branch Working Paper, Statistics Canada.
- HILLMER, S.C., and TRABELSI, A. (1987). Benchmarking of economic time series. *Journal of American Statistical Association*, 82, 1064-1071.
- LANIEL, N., and FYFE, K. (1989). Benchmarking of economic time series. Methodology Branch Working Paper, Statistics Canada.
- LANIEL, N., and FYFE, K. (1990). Benchmarking of economic time series. *Survey Methodology*, 16, 271-277.
- LANIEL, N., and MIAN, I.U.H. (1991). Maximum likelihood estimation for the constant bias model with mixed benchmarks. Methodology Branch Working Paper, Statistics Canada.
- MCLEOD, I. (1975). Derivation of the theoretical autocovariance function of autoregressive-moving average time series. *Applied Statistics*, 24, 255-256.
- MIAN, I.U.H., and LANIEL, N. (1991). Maximum likelihood estimation for the constant bias benchmarking model. Methodology Branch Working Paper, Statistics Canada.
- RAO, C.R. (1973). *Linear Statistical Inference and Its Applications*, (2nd Ed.). New York: John Wiley.

Optimum Two-Stage Sample Design for Ratio Estimators: Application to Quality Control – 1990 French Census

JEAN-CLAUDE DEVILLE¹

ABSTRACT

This study is based on the use of superpopulation models to anticipate, before data collection, the variance of a measure by ratio sampling. The method, based on models that are both simple and fairly realistic, produces expressions of varying complexity and then optimizes them, in some cases rigorously, in others approximately. The solution to the final problem discussed points up a rarely considered factor in sample design optimization: the cost related to collecting individual information.

KEY WORDS: Census quality control; Superpopulation model; Two-stage sample design optimization; Multiple objective survey.

1. INTRODUCTION

The survey method used for quality control of French census data pointed up a number of new and interesting problems, three of which are dealt with in this paper. After discussing them in general terms, we describe their specific application to the census.

In all cases, the problem is one of optimizing a two-stage survey in which the primary units are census collection districts. Units are selected using an index k that varies in a population U of districts and is, in concrete terms, a processing unit of the census forms collected.

The first problem is that of estimating the frequency of a characteristic in the population of forms (the fact of containing an error). Keeping in mind the accuracy defined for this estimate, an attempt is made to minimize survey cost with a cost function in the form

$$C_T = mC_o + nC_1, \quad (1.1)$$

where m is the number of primary units (districts) sampled, C_o the cost of processing one PU, n the number of final units (forms) sampled and C_1 the cost of processing one final unit. The problem is fairly common when a mean is to be estimated (see for example W. Cochran (1977)). Our solution is more complete as it takes into account the great variability in primary unit size.

The second, more unique, problem is also more significant. The final population (*i.e.* the forms) is made up of G separate groups ($g = 1$ to G). We are looking for an estimate of the frequency of occurrence of a characteristic in each group, with an accuracy defined for each one. The constraint resides in the fact that, because the primary units are common to all groups, sampling within one PU affects all groups.

The objective is to minimize survey cost, which is expressed as

$$C_T = mC_o + \sum_{g=1}^G n_g C_g, \quad (1.2)$$

where n_g is the total number of final units in group g and C_g the cost of processing one final unit in group g . In practice the groups are made up of the different types of census forms.

The third problem is related to coding control. We do have an *a priori* measure of the difficulty of coding each form. Formally, therefore, we have, at the level of each individual i in the population, a quantitative variable X_i , such that the probability (within a meaning to be defined) of the individual having the characteristic to be measured is approximately proportional to X_i . We are seeking to use this information to minimize the cost of control (measurement of the frequency of the "coding error" characteristic) subject to a defined survey accuracy.

In each case, plausible and simple superpopulation models allow us to evaluate the anticipated variance of the survey. In a manner of speaking, this is an almost standard illustration of model assisted survey sampling as described in Särndal, Swensson, Wretman (1992).

2. OPTIMUM ESTIMATE OF THE PROPORTION OF RECORDS CONTAINING ERRORS TWO-STAGE SAMPLE DESIGN

Each primary unit k (district) has a known number N_k of individuals (forms). Of this number, D_k display the characteristic of interest (*i.e.* contain an error). The aim is to estimate:

¹ Jean-Claude Deville, Chef de la Division des Méthodes Statistiques et Sondages, Institut National de la Statistique et des Études Économiques, 18, boul. Adolphe Pinard, 75675 Paris, CEDEX 14.

$$P = \sum_U D_k / \sum_U N_k.$$

The survey is done by drawing a sample s of primary units (PU), with π_k , the probability of inclusion in the first order and $\pi_{k\ell}$ in the second order, to be determined. Subsequently, if primary unit k is drawn in s , n_k individuals drawn by simple random sampling without replacement are checked; d_k denotes the number of forms containing errors that will be found.

Estimator \hat{P}_k of $P_k = D_k/N_k$ is expressed $\hat{P}_k = d_k/n_k$ and $\hat{D}_k = N_k \hat{P}_k$ gives an unbiased estimate of D_k . The estimator of P is expressed

$$\hat{P} = \frac{\sum_s \frac{\hat{D}_k}{\pi_k}}{\sum_s \frac{\hat{N}_k}{\pi_k}}. \quad (2.1)$$

This is the ratio of the unbiased estimators of D and N , the total number of forms. Although this number is known, estimator (4.1) is obviously more accurate than $1/N \sum_s \hat{D}_k / \pi_k$.

We have

$$\text{Var}(\hat{P}) = E \text{Var}(\hat{P} | s) + \text{Var} E(\hat{P} | s). \quad (2.2)$$

Now

$$\text{Var}(\hat{P} | s) = \hat{N}^{-2} \sum_s \frac{N_k^2}{\pi_k^2} \frac{P_k(1-P_k)N_k}{N_k-1} \left(\frac{1}{n_k} - \frac{1}{N_k} \right)$$

where

$$\hat{N} = \sum_s \frac{N_k}{\pi_k}.$$

Hence

$$E \text{Var}(\hat{P} | s) \approx N^{-2} \sum_U \frac{N_k^2}{\pi_k} \frac{P_k(1-P_k)N_k}{N_k-1} \left(\frac{1}{n_k} - \frac{1}{N_k} \right). \quad (2.3)$$

Furthermore,

$$E(\hat{P} | s) = \frac{\sum_s \frac{D_k}{\pi_k}}{\sum_s \frac{N_k}{\pi_k}}.$$

The variance of this value is obtained by linearization following introduction of variable $Z_k = D_k - PN_k = N_k(P_k - P)$.

We obtain

$$\text{Var} E(\hat{P} | s) \approx N^{-2} \text{Var} \left(\sum_s \frac{Z_k}{\pi_k} \right).$$

Taking into account that $\sum_U Z_k = 0$:

$$\text{Var} E(\hat{P} | s) = N^{-2} \left(\sum_k \frac{Z_k^2}{\pi_k} + \sum_{k \neq \ell} \frac{Z_k Z_\ell}{\pi_k \pi_\ell} \pi_{k\ell} \right). \quad (2.4)$$

The sum of (2.3) and (2.4) gives us the variance of estimator (2.1).

2.1 Introduction of a Model

Not only is the variance of \hat{P} difficult to manipulate, it contains unknown parameters. The problem may be circumvented by formulating the hypotheses required to produce a superpopulation model. It is assumed below that the parameters of this model may be estimated from the results of a preliminary test covering a very small portion of the population. In the model, expectation is denoted by E_ξ (variance by Var_ξ) and all the random variables are assumed independent of the sampling process.

The model has the following specifications:

- (a) D_k has a binomial distribution (N_k, p_k) . In the model, P_k is thus an estimator of p_k .
- (b) p_k is itself random; we assume p_k to be independent and have the same distribution, with

$$E_\xi p_k = P,$$

$$\text{Var}_\xi p_k = \sigma^2$$

for any k , in particular whatever the value of N_k .

In the model, after conditioning with p_k , we obviously have

$$E_\xi(D_k | p_k) = N_k p_k,$$

$$\text{Var}_\xi(D_k | p_k) = N_k p_k(1 - p_k).$$

The anticipated variance of \hat{P} is $E_\xi \text{Var} \hat{P}$, to which we now turn our attention. For its evaluation, we denote

$$(a) E_\xi(P_k - P)^2 = E_\xi(E_\xi(P_k - p_k + p_k - P)^2 | p_k)$$

$$= \frac{P(1-P) - \sigma^2}{N_k} + \sigma^2,$$

$$(b) E_{\xi} P_k (1 - P_k) = E_{\xi} (E_{\xi} ((P_k - P_k^2) | p_k))$$

$$= E_{\xi} p_k (1 - p_k) \frac{N_k - 1}{N_k}$$

$$= (P(1 - P) - \sigma^2) \frac{N_k - 1}{N_k},$$

(c) $E_{\xi} Z_k Z_{\ell} = 0$, because of the independence of Z_k and Z_{ℓ} , clearing one extremely cumbersome term and $\pi_{k\ell}$.

When we combine all the pieces of (2.3) and (2.4), a minor algebraic miracle occurs, producing the expression

$$E_{\xi} \text{Var } \hat{P} \approx N^{-2} \sum_U \frac{N_k^2}{\pi_k} \left(\sigma^2 + \frac{\tau^2}{n_k} \right) \quad (2.1.1)$$

where $\tau^2 = P(1 - P) - \sigma^2$
(by nature a positive quantity)

Comment:

The algebraic miracle is easily explained if we are not seeking the variance in the sole context of sample design. It is in fact the result of a model slightly more general than the one suggested.

Suppose we wish to estimate the total $N\bar{Y} = \sum_U Y_i$ of a variable Y and suppose that, to this end, a two-stage sample is drawn: in the first stage, primary units k are drawn with π_k probability and, in the second, n_k final units are drawn by simple random sampling.

We are assuming a model in which:

$$Y_i = \bar{Y} + \alpha_k + \epsilon_i,$$

with α_k a variable linked to the PU of index k . α_k is independent, subject to the same zero expectation and has a variance σ^2 . ϵ_i is also independent, centred and has a variance τ^2 . With $\pi_k^* = \pi_k n_k / N_k$ (N_k = size of PU number k), the Horvitz-Thompson estimator of the total is $\hat{Y} = \sum Y_i / \pi_i^*$, the sum being extended to the sample. In the model, and conditionally in the sample, we have

$$\text{Var}_{\xi}(\hat{Y} | s) = \sum_s \frac{N_k^2}{\pi_k^2} \left(\sigma^2 + \frac{\tau^2}{n_k} \right).$$

For this expression, expectation is again expressed in the form of equation (2.1.1).

2.2 Search for an Optimum Sample Design

The maximum variance of \hat{P} is set by the criteria selected for quality control. As the survey is repeated for each processing unit, it is only natural to seek to minimize the expected survey cost given in (2.1.1), i.e.

$$E \sum_s (C_o + n_k C_1) = \sum_U \pi_k (C_o + n_k C_1). \quad (2.2.1)$$

The problem of optimization is expressed as:

$$\text{To minimize } \sum_U \pi_k (C_o + n_k C_1)$$

with the constraints

$$N^{-2} \sum_U \frac{N_k^2}{\pi_k} \left(\sigma^2 + \frac{\tau^2}{n_k} \right) \leq V_o$$

and for any k , $n_k \leq N_k$.

Let us now apply a Lagrange multiplier λ to the first constraint – which will obviously be saturated – and multipliers μ_k to the others. We obtain the solutions

$$C_o + n_k C_1 = \lambda \frac{N_k^2}{\pi_k^2} \left(\sigma^2 + \frac{\tau^2}{n_k} \right) \quad (2.2.2)$$

and, for any k :

$$C_1 \pi_k = \lambda \frac{N_k^2}{\pi_k} \cdot \frac{\tau^2}{n_k^2} + \mu_k \quad (2.2.3)$$

with

$$\mu_k = 0 \quad \text{if } n_k < N_k \quad \text{and} \quad \mu_k > 0 \quad \text{if } n_k = N_k.$$

For the use of Lagrange multipliers, see for example Luenberger (1973).

For all primary units in which $\mu_k = 0$ (the largest), we obtain

$$n_k = \frac{\tau}{\sigma} \left(\frac{C_o}{C_1} \right)^{1/2} = n^*. \quad (2.2.4)$$

Each primary unit receives the same allocation, which corresponds to the consistent accuracy principle. Going back to equation (2.2.3), we observe that, again for these primary units, the probability of inclusion π_k must be proportional to size N_k , i.e.

$$\pi_k = \lambda^{1/2} C_1^{-1/2} \frac{\tau}{n^*} N_k. \quad (2.2.5)$$

This is the standard proof of a self-weighting one-stage survey in which the first stage is drawn with probabilities proportional to a measure of size. (See for example Cochran 1977).

Since n_k is independent of N_k , it is impossible to have $n_k = N_k$ or $\mu_k > 0$ unless $N_k \leq n^*$. Equation (2.2.2) gives us the probability of inclusion to within one factor:

$$\pi_k = \lambda^{1/2} N_k \left(\frac{\sigma^2 + \tau^2/N_k}{C_o + C_1 N_k} \right)^{1/2} = \lambda^{1/2} N_k^{1/2} \left(\frac{N_k \sigma^2 + \tau^2}{N_k C_1 + C_o} \right)^{1/2}. \quad (2.2.6)$$

Relations (2.2.5), valid if $N_k \geq n^*$, and (2.2.6) valid if $N_k \leq n^*$, establish that π_k is proportional to a known variable $T_k = f(N_k)$, for which the graph is given in Figure 1.

To fully define the survey, the number m of primary units to be drawn must still be set. $T = \sum_U T_k$ is also a known quantity.

If we restrict ourselves to fixed size sampling, we have $\pi_k = m T_k / T$. m may be determined by importing this value into the variance constraint, *i.e.*

$$N^2 V_o m = T \sum_U \frac{N_k^2}{T_k} (\sigma^2 + \tau^2/n_k).$$

If, as a first approximation, assuming $T_k = N_k$, we obtain the simplified form:

$$m V_o = \sigma^2 + \tau^2/n^*.$$

We now have a full solution to the problem.

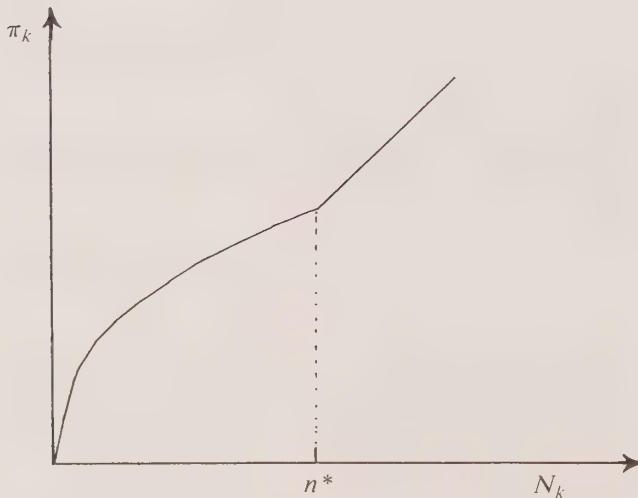


Figure 1. Graph of π_k as a function of N_k

3. OPTIMUM ESTIMATE FOR A TWO-STAGE SURVEY IN WHICH THE PRIMARY UNITS ARE STRATIFIED

The harsh facts of the situation complicate the problem somewhat: because a number of types of forms must be controlled separately, a fairly general problem, described below, arises.

For each primary unit (a district in a processing unit) we know the population N_{kg} of secondary units belonging to G groups. The “population” of PU number k is $N_{k+} = \sum_g N_{kg}$; that of group g is $N_{+g} = \sum_k N_{kg}$. As described above, we are looking for the probability of inclusion π_k with which to sample PU number k , the number of PUs to be drawn and the allocation n_{kg} of the sample among the various groups in PU k , knowing that these n_{kg} units are drawn by simple random sampling from among the N_{kg} units available.

3.1 Search for an Optimum Model Assisted Design

In each group, we postulate a model identical to the one formulated in section (2.1) (or the more general form described in the comment on that section).

For $g = 1$ to G , we have therefore:

$$v_g = E_{\xi} \text{Var}(\hat{P}_g) = N_{+g}^{-2} \sum_U \frac{N_{kg}^2}{\pi_k} (\sigma_g^2 + \tau_g^2/n_{kg}). \quad (3.1.1)$$

The cost function is expressed in the general form (1.2). We are seeking to minimize the expected survey cost

$$C_T = \sum_U \pi_k \left(C_o + \sum_g n_{kg} C_g \right), \quad (3.1.2)$$

under constraints $V_g \leq \mathfrak{V}_g$, where quantities \mathfrak{V}_g are externally fixed (*e.g.* quality of data to be obtained, tightness of control).

In this form, the problem can prove fairly complex. We write a general form of a Lagrange multiplier:

$$L = \lambda C_T + \sum_g \lambda_g V_g.$$

The problem sets $\lambda = 1$, λ_g being multipliers to be determined. In a simple variant, values are set for λ_g : we wish to minimize a given linear combination of variances under a cost constraint. In all the hypotheses, by differentiation with respect to n_{kg} (considered a real variable), we obtain

$$\lambda \pi_k^2 C_g = \lambda_g N_{+g}^{-2} N_{kg}^2 \tau_g^2 / n_{kg}^2. \quad (3.1.3)$$

π_k being for the moment to be defined to within one factor, we may write

$$\pi_k n_{kg} = \left(\frac{\lambda_g}{C_g} \right)^{1/2} \tau_g \frac{N_{kg}}{N_{+g}}. \quad (3.1.4)$$

By summing over k , we deduce that

$$E n_{+g} = \sum_U \pi_k n_{kg} = \left(\frac{\lambda_g}{C_g} \right)^{1/2} \tau_g. \quad (3.1.5)$$

The total size of the sample in each group is thus directly linked to multiplier λ_g .

Differentiation of the Lagrange multiplier with respect to π_k gives us new relations which, when combined with (3.1.4), are miraculously simplified to give

$$C_o = \sum_g C_g \left(\frac{\sigma_g}{\tau_g} \right)^2 n_{kg}^2, \quad (3.1.6)$$

or, if we introduce the numbers

$$n_g^* = \left(\frac{C_o}{C_g} \right)^{1/2} \frac{\tau_g}{\sigma_g},$$

we write

$$\sum_g \left(\frac{n_{kg}}{n_g^*} \right)^2 = 1. \quad (3.1.7)$$

As may be seen in equation (2.2.4), n_g^* is the number of secondary units to be drawn per PU if there is a single group; n_{kg} is always less than n_g^* .

From (2.1.4), (3.1.5) and (3.1.7) we obtain the relations:

$$\pi_k^2 = \frac{1}{C_o} \sum_g \lambda_g \sigma_g^2 \left(\frac{N_{kg}}{N_{+g}} \right)^2. \quad (3.1.8)$$

Thus, π_k is proportional to T_k such that $T_k^2 = \sum_g \lambda_g \sigma_g^2 N_{kg}^2 / N_{+g}^2$, which appears to be a satisfactory measure of size. The relations (3.1.4) show that, if k is fixed, n_{kg} is proportional to $n_g^* \lambda_g^{1/2} \sigma_g N_{kg} / N_{+g}$; taking into account (3.1.7), we obtain

$$n_{kg} = n_g^* \lambda_g^{1/2} \sigma_g \frac{N_{kg}}{N_{+g}} T_k^{-1}. \quad (3.1.9)$$

3.2 Explicit Solutions to Two Specific Cases

(a) If λ_g were known, *i.e.* if $\sum_g \lambda_g v_g$ were minimized under a cost constraint, then (3.1.2) and (3.1.9) could be used to calculate T_k . By transferring

$$\pi_k = m T_k / T \left(T = \sum_U T_k, m \text{ number of primary units to be drawn} \right)$$

to budget constraint $C_T \leq C_T^*$, we find that

$$C_T^* = \frac{m}{T} \sum_U \left(C_o T_k + \sum_g C_g n_g^* \lambda_g^{1/2} \sigma_g \frac{N_{kg}}{N_{+g}} \right) \quad i.e.$$

$$m = C_T^* \left(C_o + \sum_g C_g n_g^* \cdot \frac{\lambda_g^{1/2} \sigma_g}{T} \right).$$

If a single λ_g is not equal to zero, it is fairly easy to check that the result is the one given at the end of section (2.2).

(b) The initial problem ($\min C_T$ under $V_g \leq \vartheta_g$) is resolved fairly easily in two specific cases.

b1 – *Maximum dispersion* of the groups. For any PU k , we have $N_{kg} = N_{k+}$ for a given k . The problem is broken down into G separate problems, each being of the type examined in section 2.

b2 – *Minimum dispersion*. The distribution is the same in all the PUs; in other words, for any k and any g , we have

$$N_{kg} = N_{k+} \frac{N_{+g}}{N} \quad \text{with} \quad \left(N = \sum_g N_{+g} \right),$$

T_k is then proportional to N_{k+} , and n_{kg} is quantity $n_g^* u_g$ independent of k .

With $\pi_k = m N_{k+} / N$, we obtain by writing $V_g = \vartheta_g$:

$$m \vartheta_g = \sigma_g^2 + \tau_g^2 / n_g^* u_g$$

i.e.

$$m = \frac{\sigma_g^2}{\vartheta_g} + u_g^{-1} \frac{\tau_g^2}{n_g^* \vartheta_g}.$$

Thus we obtain G-1 linear relations between the u_g^{-1} , in principle permitting full resolution of the problem, knowing that the sum of u_g^2 is equal to 1.

3.3 A Numerical Algorithm for Determining the Optimum Solution to the General Case

An iterative numerical resolution of the problem may be achieved as follows.

Step 1: An approximate sample allocation is set in each group (n_{+g} units in group g). The process may be facilitated by using the approximate solution based on the hypotheses in point (a) or point (b).

Step 2: The value of λ_g is determined from relations (3.1.5):

$$\lambda_g = C_g n_{+g}^2 / \tau_g^2.$$

Step 3: π_k is determined from relations (3.1.8). Specifically, the sum of π_k sets the number of PUs to be drawn.

Step 4: n_{kg} is determined from relations (3.1.4). Subsequent iteration is possible by returning to step 2, in the expectation that the algorithm will converge toward the optimization solution.

Comment: The probability of drawing a type g unit is

$$\pi_k n_{kg} / N_{kg} = \left(\frac{\lambda_g}{C_g} \right)^{1/2} \tau_g / N_{+g}.$$

Because it does not depend on primary unit k , it is the same for each unit in a given group g (equal probability survey). Size n_{+g} , or at least its mathematical expectation, may be deduced from the sample in group g . In practice, sample size is sometimes set arbitrarily: this entails determining λ_g or, implicitly, variances τ_g^2 . This is another fairly common result.

4. OPTIMUM ESTIMATE ASSISTED BY A MEASURE OF THE DIFFICULTY OF CODING A RECORD

The task is to estimate the proportion of forms containing a coding error in universe U of all forms coded in a given week by one regional branch. The problem is identified by the following characteristic: because all IFs are precoded, it is possible, using information drawn from the trial census, to attribute to each one a positive numerical variable X_i representing its “difficulty”. This variable is calibrated in such a way that Y_i (equal to 1 if there is an error and 0 if there is not) has an “expectation” proportional to X_i .

The same cost control considerations suggest a two-stage survey.

- In the first stage of the survey, we draw a sample s_1 of districts k (primary units), with π_k unequal probabilities to be determined. $\pi_{k\ell}$ denotes the probability of inclusion, double in value in this instance.
- In the second stage of the survey, a sample s_k of final units (forms) in primary unit sample k is drawn. $\pi_{i|k}$ denotes the probability of inclusion of the unit in primary unit k , $\pi_{ij|k}$ the probability of inclusion of the pair (i, j) in the primary unit; and $s = U_{k \in s_1} s_k$, the sample of final units.

$X_k = \sum_{i \in k} X_i$ denotes the total of X_i in primary unit k , $X = \sum_{k \in U_o} X_k = \sum_U X_i$ and similar notations are used for all the variables. (U_o denotes the population of primary units – districts, U the population of final units – forms).

The aim is to estimate a quantity in the form $R = \sum_U Y_i / \sum_U W_i$ where W_i is a known variable for each form. This may be $W_i = 1$ or $W_i = X_i$, whichever measure of the error rate seems the more satisfactory.

4.1 Selection of Estimator and Variance

- (a) For primary unit k , the total Y_k of the Y_i for $i \in k$ is commonly estimated by the ratio

$$\hat{Y}_k = X_k \left(\sum_{s_k} Y_i / \pi_{i|k} \right) / \left(\sum_{s_k} X_i / \pi_{i|k} \right) = X_k \hat{a}_k$$

where \hat{a}_k estimates $a_k = Y_k / X_k$ with a slight bias.

- (b) To estimate ratio Y/X , we use

$$\hat{a} = \frac{\sum_{s_1} \frac{\hat{Y}_k}{\pi_k}}{\sum_{s_1} \frac{X_k}{\pi_k}} = \frac{\sum_{s_1} \hat{a}_k \frac{X_k}{\pi_k}}{\sum_{s_1} \frac{X_k}{\pi_k}}.$$

- (c) If we wish to estimate R , we note that

$$R = \frac{Y}{X} \cdot \frac{X}{W},$$

where X and W are known totals (e.g. total difficulty, total number of forms). As variable X_i was selected for its good correlation with Y_i , an *a priori* valuable estimator of R is

$$\hat{R} = \hat{a} \frac{X}{W}$$

and the only real question concerns the estimate of $a = \sum_k a_k X_k / X$.

- (d) we have

$$\text{Var}(\hat{a}) = \text{Var} E(\hat{a} | s_1) + E \text{Var}(\hat{a} | s_1).$$

For the first term, taking into account the fact that \hat{a}_k is an approximate unbiased estimator of a_k , we may write

$$\begin{aligned} \text{Var} E(\hat{a} | s_1) &\approx \frac{1}{X^2} \text{Var} \left(\sum_{s_1} \frac{(a_k - a) X_k}{\pi_k} \right) \\ &= \frac{1}{X^2} \left(\sum_k \frac{(a_k - a)^2 X_k^2}{\pi_k^2} \right. \\ &\quad \left. + \sum_{k \neq \ell} (a_k - a)(a_\ell - a) \frac{X_k X_\ell \pi_{k\ell}}{\pi_k \pi_\ell} \right). \quad (4.1.1) \end{aligned}$$

For the second term, *conditional on* s_1 , we have

$$\text{Var} \left(\frac{\sum_{s_1} \hat{a}_k \frac{X_k}{\pi_k}}{\sum_{s_1} \frac{X_k}{\pi_k}} \right) = \left(\sum_{s_1} \frac{X_k}{\pi_k} \right)^{-2} \cdot \sum_{s_1} \text{Var}(\hat{a}_k) \frac{X_k^2}{\pi_k^2}.$$

For this quantity, the expectation is approximately

$$X^{-2} \sum_k E \text{Var}(\hat{a}_k | s_1) \frac{X_k^2}{\pi_k}, \quad (4.1.2)$$

with

$$\begin{aligned} \text{Var}(\hat{a}_k | s_1) &= \text{Var} \frac{\sum_{s_k} \frac{Y_i}{\pi_{i|k}}}{\sum_{s_k} \frac{X_i}{\pi_{i|k}}} \approx \frac{1}{X_k^2} \text{Var} \sum_{s_k} \frac{Y_i - a_k X_i}{\pi_{i|k}} \\ &= \frac{1}{X_k^2} \left(\sum_{i \in k} \frac{(Y_i - a_k X_i)^2}{\pi_{i|k}} \right. \\ &\quad \left. + \sum_{k \neq l} \frac{(Y_i - a_k X_i)(Y_j - a_l X_j) \pi_{ij|k}}{\pi_{i|k} \pi_{j|k}} \right). \end{aligned}$$

As in the preceding sections, we arrive at formulae that are complex and, in the final analysis, unusable. A model will simplify things somewhat.

4.2 Introduction of a Model

The model has the same structure as those used previously:

- (a) a_k is an independent random variable with the same expectation and the same variance:

$$E_{\xi} a_k = a \quad \text{Var}_{\xi} a_k = \sigma^2.$$

The variance takes into account operator influence, which we make no attempt to isolate, and also such factors as day of the week, time of day, day of the month *etc.* . . .

- (b) Conditional on a_k , Y_i in primary unit k is an independent Bernoulli variable with $E_{\xi}(Y_i | k) = a_k X_i$

$$\text{Var}_{\xi}(Y_i | k) = a_k X_i - a_k^2 X_i^2.$$

Comment:

Variable X_i , which has no actual concrete meaning, is defined to within one factor of scale. Conversely aX_i and σX_i , being probabilities, have an invariant physical interpretation. In what follows, one must always keep in mind that the results are invariant if X_i is multiplied by an arbitrary factor, on condition that a and σ are divided by the same factor. $\text{Var}(\hat{a})$ in particular has no concrete meaning; $\text{Var}(\hat{a}X)$ is an exception.

As before, we examine anticipated variance, expectation under the model of the sum of (4.1.1) and (4.1.2).

For the first term, the expectation of the cross products is of course zero. The expectation under the model for this term is thus:

$$X^{-2} \sigma^2 \sum_k \frac{X_k^2}{\pi_k}.$$

For the second term, we find (in light of the definitions given in 4.2.a and 4.2.b)

$$\begin{aligned} X^{-2} \sum_k \frac{X_k^2}{\pi_k} \cdot \frac{1}{X_k^2} \sum_i E_{\xi} \frac{(a_k X_i - a_k^2 X_i^2)}{\pi_{i|k}} \\ = X^{-2} \sum_k \frac{1}{\pi_k} \sum_i \frac{a X_i - (a^2 + \sigma^2) X_i^2}{\pi_{i|k}}. \end{aligned}$$

Therefore, overall

$$\begin{aligned} E_{\xi} \text{Var}(\hat{a}X) &= \sigma^2 \sum_{k \in U_0} \frac{X_k^2}{\pi_k} \\ &\quad + \sum_{k \in U_0} \frac{1}{\pi_k} \sum_{i \in k} \frac{a X_i - (a^2 + \sigma^2) X_i^2}{\pi_{i|k}}. \end{aligned}$$

No algebraic miracle occurs here. *For simplification*, we assume that $(a^2 + \sigma^2) X_i^2$ is negligible in the face of $a X_i$. Numerically, we may expect $a X_i = 2$ to 5×10^{-2} and $(a^2 + \sigma^2) X_i^2 = 3$ to 30×10^{-4} ; whence the approximation

$$E_{\xi} \text{Var}(\hat{a}X) \approx \sigma^2 \sum_{k \in U_0} \frac{X_k^2}{\pi_k} + a \sum_{k \in U_0} \frac{1}{\pi_k} \sum_{i \in k} \frac{X_i}{\pi_{i|k}}.$$

4.3 Sample Design Optimization

We use the following cost function:

$$C = \sum_{s_1} (C_0 + C_1 n_k).$$

Here, $n_k = \sum_{i \in k} \pi_i |k|$ is the size of the sample drawn in district k (supposedly set at fixed size s_1). Its expectation is

$$C_T = \sum_{k \in U_o} \pi_k (C_o + C_1 n_k).$$

Let

$$\pi_{i|k} = n_k P_i \left(\text{with } \sum_{i \in k} P_i = 1 \right) \quad \text{and} \quad Q_k = \pi_k n_k.$$

The problem of optimization is now

$$\begin{aligned} \text{Min: } & C_o \sum_k \pi_k + C_1 \sum_k Q_k \\ \text{under: } & \sigma^2 \sum_k \frac{X_k^2}{\pi_k} + a \sum_k \frac{1}{Q_k} \sum_{i \in k} \frac{X_i}{P_i} \leq \gamma_o. \end{aligned}$$

In this form, we are pleased to observe that the terms in $\sum_i X_i / P_i$ may be minimized independently of the other terms. In other words, n_k has no impact on this term. Leaving optimization of the second stage of the survey until later, S_k^{*2} denotes the optimized value of $\sum_i X_i / P_i$.

With a Lagrange multiplier λ , by differentiation with respect to π_k and Q_k , we obtain

$$\begin{aligned} *C_o = \lambda \sigma^2 \frac{X_k^2}{\pi_k^2} \quad \text{i.e.} \quad & \pi_k \text{ proportional to } X_k \quad (4.3.1) \\ *C_1 = \lambda a \frac{S_k^{*2}}{Q_k^2} \quad \text{whence} \quad & n_k = \left(\frac{C_o}{C_1} \right)^{1/2} \frac{a^{1/2}}{\sigma} \frac{S_k^*}{X_k}. \quad (4.3.2) \end{aligned}$$

Specifically, the primary units are drawn with probabilities proportional to total difficulty, a standard resolution (see for example Särndal, Swensson, Wretman, 1992, Chapter 12).

We now move on to sub-district sampling (second stage of survey).

Beginning with a simple, straightforward case, forms are drawn one by one. Minimization produces P_i proportional to $\sqrt{X_i}$. A simple calculation shows that $S_k^* = \sum_{i \in k} \sqrt{X_i}$. We can now calculate n_k using (4.3.2), and our problem is fully resolved.

In practice, things are more complicated. For fairly obvious reasons, only forms for entire households are selected. In other words, the second stage of the survey is a *cluster* survey. The values of P_i are the same (*i.e.* P_m) for all the members of a given cluster (household) m .

Let X_m be the sum of X_i individuals i in household m . The problem is to minimize $\sum X_m / P_m$ under $\sum n_m P_m = 1$, with n_m the size of household m . We easily reach solution

$$P_m = \sqrt{X_m} / \sum n_m \sqrt{X_m},$$

with $\bar{X}_m = X_m / n_m$, *mean difficulty of forms IF in household m* . From this we determine $S_k^* = \sum n_m \sqrt{\bar{X}_m}$.

This solution enables us to determine the number n_k of *final units* to be drawn using (4.3.2). However, the number of clusters (*households*) has not been determined: this snag was predictable. In fact, the cost function does not imply this constraint. To obtain the number m_k of clusters to be drawn, we arrange matters so that the expectation of the number of final units is equal to n_k . Thus,

$$m_k \left(\sum n_m \sqrt{\bar{X}_m} \right) / \sum \sqrt{\bar{X}_m}$$

whence

$$m_k = n_k \frac{\sum \sqrt{\bar{X}_m}}{\sum n_m \sqrt{\bar{X}_m}}.$$

Taking into account (4.3.2), we also have

$$m_k = \left(\frac{C_o}{C_1} \right)^{1/2} \frac{a^{1/2}}{\sigma} \frac{\sum \sqrt{\bar{X}_m}}{X_k}$$

and the probability a given household being drawn is thus

$$\frac{m_k \sqrt{\bar{X}_m}}{\sum \sqrt{\bar{X}_m}}.$$

Following a number of algebraic manipulations, the value of the optimum variance is found to be:

$$E_{\xi} \text{Var}(\hat{a}X)_{\text{OPT}} = \frac{(\sigma X)^2}{m} \left(1 + \frac{a}{\sigma} \frac{a^{-1/2} S^*}{X} \left(\frac{C_1}{C_o} \right)^{1/2} \right).$$

This form respects the homogeneous character of the different factors. In particular, we have $a^{-1/2} S^* / X = a^{1/2} S^* / aX$: the denominator may be interpreted as total number of errors in a lot; the numerator is homogeneous for a given size.

We now have a full solution to the problem.

Comment 1:

In both cases discussed, S_k^* is multiplied by $C^{1/2}$ if X_i is multiplied by C . The formula that gives n_k is thus invariant on the scale of measurement.

Comment 2:

The solution that entails drawing clusters favours small clusters made up of final units with a high index of difficulty.

Comment 3:

As in preceding sections, we determine the probability of single selection, but not the probability of dual selection. Therefore the algorithm for the draw, which sets the latter, has no influence. This is quite common, keeping in mind that the complementary data used to optimize the draw determines π_k and $\pi_{i|k}$ but have no influence on dual probabilities.

5. APPLICATIONS TO CONTROL BY SURVEY OF THE QUALITY OF THE 1990 FRENCH CENSUS

5.1 Problem of Data Capture Control

The sampling techniques described in sections 2 and 3 were designed to control data capture for the 1990 Census. A brief description of the operation would enhance understanding of the nature of the statistical problems involved.

The basic collection unit is the district, which corresponds, in a city, to a block of houses and, in the country, to a village or group of hamlets. It covers a population that ranges from zero inhabitants to approximately 2,000 (the mean values are 150 dwellings and approximately 350 inhabitants).

When collection is completed and the results are audited, the various census forms (specifically individual forms (IF) and dwelling forms (DF)) are meticulously counted for each district. The summary data for a district are computerized; the forms themselves, collated into district files, are forwarded to data capture.

Groups of districts comprising approximately 100,000 dwellings are constructed. The processing units (PU) are processed for INSEE by contractors. INSEE, the "client" in terms of control theory, monitors the quality of each contractor's work by sampling a specific number of forms in each PU.

The aim of the survey described in paragraph 2 is to estimate, to an accuracy (standard deviation) of one point, the proportion of forms containing an error in each PU. The maximum proportion of forms containing an error cannot exceed 4%. A trial census covering approximately 400 districts allows for an estimate of the values of the two model parameters. We find:

$$\sigma^2 \approx P^2 \approx 14.10^{-4}$$

$$\tau^2 \approx P \approx 4.10^{-2}.$$

Cost function (1.1) is assessed in terms of working time. Based on on-site control measures, 5 minutes is the estimate of the time required to process one district folder (from the time it is taken from the shelf to the time it is returned there) and 30 seconds the estimate of the time required to process one IF. With the numerical data, design optimization based on the hypotheses in section 1 allows for control of 40 districts per processing lot and 16 forms per district.

After discussing the solution with the team responsible for the census, it emerged that two types of documents (individual forms (IF) and dwelling forms (DF)) were to be controlled. The first approximation had taken no account of the latter, which are less likely to contain errors and take only about half as long to code as IFs. However, some districts (e.g. a commune with a thriving tourist industry) contain a large majority of secondary dwellings, and so produce many DFs but very few IFs. Because the situation required in-depth study, the theory given in section 3 was developed.

In the case of the census, the number of groups G is equal to 2 ($g = 1$ for the IFs and $g = 2$ for the DFs). The numerical data for the two groups are:

$$\begin{aligned} \cdot P_1 &= 0,04 & \sigma_1 &= P_1 & \tau_1^2 &= P_1(1 - P_1) \\ & & & & & - \sigma_1^2 = P_1 - 2P_1^2, \end{aligned}$$

$$\cdot P_2 = 0,01 \quad \sigma_2 = P_2 \quad \tau_2^2 = P_2 - P_2^2,$$

$$\cdot \varpi_1 = (0,0075)^2 \quad \varpi_2 = (0,0150)^2.$$

For the cost function, we selected $C_o = 5$ minutes, $C_1 = 0.5$ minute and $C_2 = 0.25$ minute. Optimization of the problem according to the hypotheses in section 3.2.b entailed examining 73 districts per processing unit. In practical terms, it meant processing 15 individual forms (and related DFs) for each district. For the districts that produce fewer than 15 IFs, all IFs were processed. For districts with zero IFs, 4 DFs were processed (if this number was less than the number of DFs in the district).

Comment:

The method described in part 2 seems to have a fairly broad field of application. One example: it was used to sample the 1992 French survey on migration of foreign nationals. For population centres with under 20,000 inhabitants, the sample was drawn in two stages. The first stage of the survey covered the 90 departments in which this type of population centre occurs. The foreign population (based on the census) was divided into 8 nationality groups, for which equally accurate indicators had to be found.

5.2 Problems Related to Coding

The second step in data preparation is known as operation COLIBRI (Codification en Ligne des Bulletins du Recensement des Individus). The operators in the regional branches of INSEE receive forms classified by district and code them for the 25% survey.

In practice, each operator works at a monitor that displays the identifier of the next dwelling to be included in the 25% sample, for which all IFs must be coded.

Coding quality is also controlled by survey. The control unit is all the work done in one week in a regional branch. The entire operation takes a little over one year in the 22 regional branches, and entails more than 1,000 surveys. The household is the unit to be controlled (*i.e.* all the IFs in a household drawn for inclusion in the control sample). The objective is to estimate the proportion of forms containing an error. This is done by automatic detection of forms in which there is a no match situation. The number of errors is determined by reconciliation. The control theory is discussed in section 4 of this paper. The index of difficulty of the forms was developed from the data captured for a study based on the previous census and by test. The procedure and results related to these control measures are described in detail in G. Badeyan (1992).

The practical and numerical application of the theory rests on hypotheses concerning the orders of magnitude of the different parameters (which requires linking them to a simple physical interpretation). In the census preparation phase, without accurate prior measurement, we used the values $\sigma/a = 0.5$ and $C_1/C_0 = 0.1$.

Pursuant to a number of hypotheses concerning the other parameters, and after discussing the matter with experts, it was decided that the control would cover 50 districts, with approximately 20 IFs controlled in each one (by region and by week). Since model parameters can be re-estimated at any stage in the process, the initial order of magnitude can obviously be adjusted as the survey proceeds.

Final Comment:

The problem produces somewhat surprising results that are worthy of consideration.

In the first instance, as we assumed it would be possible to separate each form, the forms were drawn with a probability proportional to individual difficulty. We assumed, to some extent, that the cost of using individual information was zero.

In the second instance, the actual control process, it was assumed that cost was infinite and the only information

with negligible cost was the information related to an entire household. The solution shows that the probability of drawing an individual (IF) as a function of the mean difficulty of coding the forms for the entire household of which the individual is a member.

The same phenomenon occurs in the district draw. If it is possible to separate the IFs, they are drawn with probabilities proportional to total difficulty; within a district, the difficult IF has a greater probability of selection. Conversely, suppose we are unable to separate IFs within a district. This will be the case, for example, if the designation of IFs to be controlled cannot be implemented in real time because of inadequate processing facilities. Districts would then be selected in proportion to mean difficulty: within a district, it would be necessary to proceed by simple random sampling.

In the first instance, the survey gives precedence to large districts, from which difficult IFs tend to be drawn. In the second instance, precedence is given to small difficult districts, from which forms are selected with equal probability. *In both instances*, we are seeking to increase the probability of surveying difficult IFs. The difference resides simply in the possibility (*i.e.* the cost) of collecting information when we need it.

ACKNOWLEDGEMENTS

The author would like to thank the Editor, the Associate Editor and the Referee for their extremely positive comments. He would also like to thank Claude Thelot, some of whose comments have been incorporated into this paper and Gérard Badeyan, who introduced the methods discussed here at INSEE. Last, but not least, he would like to thank Françoise Hitier, without whom this paper would never have seen the light of day.

REFERENCES

- BADEYAN, G. (1992). Communication aux secondes Journées de Méthodologie Statistique, June 17 and 18, 1992, INSEE, Paris.
- COCHRAN, W. (1977). *Sampling Techniques*, (3rd Edition). New York: Wiley.
- DESABIE, J. (1965). *Théorie et Pratique des Sondages*. Paris: Dunod.
- LUENBERGER, D.G. (1973) *Introduction To linear and Non-linear Programming*. New York: Addison-Wesley.
- SÄRNDAL, C.-E., SWENSSON, B., and WRETMAN, J. (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.

Conditional Properties of Post-Stratified Estimators Under Normal Theory

ROBERT J. CASADY and RICHARD VALLIANT¹

ABSTRACT

Post-stratification is a common technique for improving precision of estimators by using data items not available at the design stage of a survey. In large, complex samples, the vector of Horvitz-Thompson estimators of survey target variables and of post-stratum population sizes will, under appropriate conditions, be approximately multivariate normal. This large sample normality leads to a new post-stratified regression estimator, which is analogous to the linear regression estimator in simple random sampling. We derive the large sample design bias and mean squared errors of this new estimator, the standard post-stratified estimator, the Horvitz-Thompson estimator, and a ratio estimator. We use both real and artificial populations to study empirically the conditional and unconditional properties of the estimators in multistage sampling.

KEY WORDS: Asymptotic normality; Regression estimator; Defective frames; Ratio estimator; Horvitz-Thompson estimator.

1. INTRODUCTION

1.1 Background

A major thrust in sampling theory in the last twenty years has been to devise ways of restricting the set of samples used for inference. In a purely design-based approach, as described in Hansen, Madow, and Tepping (1983), no such restrictions are imposed. Statistical properties are calculated by averaging over the set of all samples that might have been selected using a particular design. Although it is generally conceded that some type of design-based, conditional inference is desirable (Fuller 1981, Rao 1985, Hidiroglou and Särndal 1989), satisfactory theory has yet to be developed except in relatively simple cases. Alternative approaches are prediction theory, developed by Royall (1971) and many others, and the Bayesian approach, found in Ericson (1969), which avoid averaging over repeated samples through the use of superpopulation models. A design-based approach to conditioning was introduced by Robinson (1987) for the particular case of ratio estimates in sample surveys. Robinson applied large sample theory and approximate normality of certain statistics to produce a conditional, design-based theory for the ratio estimator.

In this paper, we extend that line of reasoning to the problem of post-stratification. Convincing arguments have been made in the past by Durbin (1969), Holt and Smith (1979) and Yates (1960) that post-stratified samples should be analyzed conditional on the sample distribution of units among the post-strata. However, as Rao (1985) has noted, the difficulties in developing an exact, design-based, finite sample theory for post-stratification in general

sample designs may be intractable. Model-based, conditional analyses of post-stratified samples are presented in Little (1991) and Valliant (1993). The alternative pursued here is design-based and uses large sample, approximate normality in a way similar to that of Robinson (1987) as a means studying conditional properties of estimators.

1.2 Basic Definitions and Notation

The **target population** is a well defined collection of elementary (or analytic) units. For many applications the elementary units are either persons or establishments. We assume the target population has been partitioned into **first stage sampling units (FSUs)**. For person based surveys the FSUs are commonly households, groups of households or even counties, while for establishment based surveys it is not uncommon that the individual establishment is an FSU. In any event, the collection of FSUs will be referred to as the **first stage sampling frame** (or just **sampling frame**). It is assumed that there are M FSUs in the sampling frame and they are labeled $1, 2, \dots, M$. We also assume that the population units can be partitioned into K "post-strata" which can be used for the purposes of estimation.

We let y represent the value of the characteristic of interest (*e.g.* weekly income, number of hours worked last week, restricted activity days in last two weeks, *etc.*) for an elementary unit. Associated with the i^{th} FSU are $2K$ real numbers:

y_{ik} = aggregate of the y values for the elementary units in the i^{th} FSU which are in the k^{th} post-stratum,
 N_{ik} = number of elementary units in the i^{th} FSU which are in the k^{th} post-stratum.

¹ Robert J. Casady and Richard Valliant, U.S. Bureau of Labor Statistics, 2 Massachusetts Ave. N.E., Washington D.C., 20212-0001.

For each post-stratum we then define

$$Y_{.k} = \sum_{i=1}^M y_{ik} = \text{aggregate of the } y \text{ values for all elementary units in the } k^{\text{th}} \text{ post-stratum,}$$

$$N_{.k} = \sum_{i=1}^M N_{ik} = \text{total number of elementary units in the } k^{\text{th}} \text{ post-stratum.}$$

In what follows we assume that the $N_{.k}$ are known fixed values. In some surveys, the $N_{.k}$ may actually be estimates themselves but our analysis is conditional on the set of $N_{.k}$ used in estimation. In the Current Population Survey in the United States, for example, each $N_{.k}$ is a population count projected from the previous decennial census using demographic methods. The population aggregate of the y values is given by $Y_{..} = \sum_{k=1}^K Y_{.k}$ and the total population size by $N_{..} = \sum_{k=1}^K N_{.k}$. In sections 1-3, we assume that the sampling frame provides “coverage” of the entire target population. In section 4, we consider the problem of a defective frame, *i.e.* one in which the coverage of the frame differs from that of the target population.

1.3 Sample Design and Basic Estimation

Suppose that the first stage sampling frame is partitioned into L strata and that a multi-stage, stratified design is used with a total sample of m FSUs. In the following, the subscript representing design strata is suppressed in order to simplify the notation. For the subsequent theory, it is unnecessary to explicitly define sampling and estimation procedures for second and higher levels of the design. However, for every sample FSU, we require estimators \hat{y}_{ik} and \hat{N}_{ik} so that $E_{2+}[\hat{y}_{ik}] = y_{ik}$ and $E_{2+}[\hat{N}_{ik}] = N_{ik}$ where the notation E_{2+} indicates the design-expectation over stages 2 and higher. Letting π_i be the probability that the i^{th} FSU is included in the sample and $w_i = 1/\pi_i$, it follows that the estimator $\hat{Y}_{.k} = \sum_{i=1}^m w_i \hat{y}_{ik}$ is unbiased for $Y_{.k}$ and the estimator $\hat{N}_{.k} = \sum_{i=1}^m w_i \hat{N}_{ik}$ is unbiased for $N_{.k}$.

1.4 An Analogue to Robinson's Asymptotic Result

Robinson (1987) studied the ratio estimator $(\bar{X}/\bar{x}_s)\bar{y}_s$ under simple random sampling with \bar{y}_s being the sample mean of a target variable y , \bar{x}_s being the sample mean of an auxiliary variable x , and \bar{X} the population mean of x . Under certain conditions (\bar{y}_s, \bar{x}_s) will be asymptotically, bivariate normal in large simple random samples. From Robinson's results it follows that the linear regression estimator $\bar{y}_s + \beta(\bar{X} - \bar{x}_s)$ is asymptotically design-unbiased conditional on \bar{x}_s . Results in this section extend that result to complex samples.

Following Krewski and Rao (1981), we can establish our asymptotic results as $L \rightarrow \infty$ within the framework of a sequence of finite populations $\{\Pi_L\}$ with L strata in Π_L . It should be understood that we implicitly assume (without formal statement) the sample design and regularity conditions as specified in Krewski and Rao and more fully developed in Rao and Wu (1985). Details of proofs add little to those in the literature and are omitted.

Converting to matrix notation, we let $\mathbf{Y} = [Y_{.1} \dots Y_{.K}]'$, $\mathbf{N} = [N_{.1} \dots N_{.K}]'$, $\hat{\mathbf{Y}} = [\hat{Y}_{.1} \dots \hat{Y}_{.K}]'$, $\hat{\mathbf{N}} = [\hat{N}_{.1} \dots \hat{N}_{.K}]'$ and $\mathbf{V} = \text{var}\{[\hat{\mathbf{Y}} \hat{\mathbf{N}}]'\}$ where $\hat{\mathbf{Y}} = (1/N_{..})\hat{\mathbf{Y}}$ and $\hat{\mathbf{N}} = (1/N_{..})\hat{\mathbf{N}}$. Note that $\hat{\mathbf{Y}}$, which uses $N_{..}$ in the denominator, is a notational convenience and does not estimate means in the post-strata. Analogous to conditions C4 and C5 of Krewski and Rao (1981), we assume that

$$\lim_{L \rightarrow \infty} \frac{Y_{.k}}{N_{.k}} = \mu_k, \quad \text{for } k = 1, 2, \dots, K, \quad (1)$$

$$\lim_{L \rightarrow \infty} \frac{N_{.k}}{N_{..}} = \phi_k > 0 \quad \text{for } k = 1, 2, \dots, K, \text{ and } (2)$$

$$\lim_{L \rightarrow \infty} m\mathbf{V} = \Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} \text{ (positive definite), } (3)$$

where Σ is partitioned in the obvious manner. Note that we have again suppressed the subscript representing design strata. Assumptions (1)–(3) simply require that certain key quantities stabilize in large populations. Condition (2), in particular, assures that no post-stratum is empty as the population size increases. We now state the following.

Result: Assume the sample design and regularity conditions specified in Krewski and Rao and that Σ_{22}^{-1} exists; then, given $\hat{\mathbf{N}}$, the conditional distribution of $\hat{\mathbf{Y}}$ is asymptotically $\mathcal{N}(\mathbf{M}_1 + \Sigma_{12}\Sigma_{22}^{-1}(\hat{\mathbf{N}} - \mathbf{M}_2), m^{-1}\mathbf{V}_c)$, where $\mathbf{V}_c = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$, $\mathbf{M}_1 = \lim_{L \rightarrow \infty} \hat{\mathbf{Y}} = [\phi_1 \mu_1 \dots \phi_K \mu_K]'$ and $\mathbf{M}_2 = \lim_{L \rightarrow \infty} \hat{\mathbf{N}} = [\phi_1 \dots \phi_K]'$.

Proof. This result is analogous to the result for $K = 1$ given by Robinson (1987) and follows directly from the fact that the random vector

$$m^{1/2} \begin{bmatrix} \hat{\mathbf{Y}} - \mathbf{M}_1 - \Sigma_{12}\Sigma_{22}^{-1}(\hat{\mathbf{N}} - \mathbf{M}_2) \\ \hat{\mathbf{N}} - \mathbf{M}_2 \end{bmatrix}$$

tends in distribution to

$$N\left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \mathbf{V}_c & \mathbf{0} \\ \mathbf{0} & \Sigma_{22} \end{bmatrix}\right).$$

Strictly, as in Robinson, we consider the conditional distribution of $\hat{\mathbf{Y}}$ for $\hat{\mathbf{N}}$ in a cell of size $\epsilon m^{-1/2}$ for small ϵ . Note that in some sample designs $\mathbf{1}'\hat{\mathbf{N}} = N_{..}$ (such as those in which a fixed number of elementary units are selected with equal probabilities) in which case Σ_{22}^{-1} does not exist; in such cases only the first $K - 1$ post-strata are considered for the purpose of conditioning.

In the next section, the asymptotic mean of $\hat{\mathbf{Y}}$ is used to motivate a linear regression estimator of the population mean of the y 's.

2. CONDITIONAL PROPERTIES OF ESTIMATORS FOR THE POPULATION MEAN

2.1 Estimators for the Population Mean

The **population mean** is, by definition,

$$\mu = \lim_{L \rightarrow \infty} (Y_{..}/N_{..}) = \lim_{L \rightarrow \infty} (\mathbf{1}' Y / \mathbf{1}' N) = \sum_{k=1}^K \phi_k \mu_k$$

where $\mathbf{1}'$ is a row vector of K ones. Note that the mean μ is not a finite population parameter but rather a limiting value. In large populations ($L \rightarrow \infty$) μ and the actual finite population mean will be arbitrarily close. Four estimators of the population mean will be considered. The first three are standard estimators found in the literature while the fourth is a new estimator motivated by the asymptotic, joint normality of \hat{Y} and \hat{N} :

(1) Horvitz-Thompson estimator

$$\hat{Y}_{HT} = \mathbf{1}' \hat{Y} / \mathbf{1}' N = \mathbf{1}' \hat{Y}.$$

(2) Ratio estimator

$$\hat{Y}_R = \mathbf{1}' \hat{Y} / \mathbf{1}' \hat{N} = \mathbf{1}' \hat{Y} / \mathbf{1}' \hat{N}.$$

(3) Post-stratified estimator

$$\hat{Y}_{PS} = N_{..}^{-1} \sum_{k=1}^K \left(\frac{N_{..k}}{\hat{N}_{..k}} \right) \hat{Y}_{..k} = \mathbf{r}' \hat{Y}$$

where

$$\mathbf{r}' = [N_{..1}/\hat{N}_{..1}, \dots, N_{..K}/\hat{N}_{..K}].$$

(4) Linear regression estimator

$$\hat{Y}_{LR} = [\mathbf{1}' (\hat{Y} - \sum_{12} \sum_{22}^{-1} (\hat{N} - \mathbf{M}_2))] .$$

The linear regression estimator is motivated by the form of the large sample mean of the conditional random variable $\hat{Y} | \hat{N}$ listed at the end of section 1.4 and is very similar to the generalized regression estimator discussed by Särndal, Swensson and Wretman (1992). The linear regression estimator (4) was also discussed in the context of calibration estimation by Rao (1992). It should be noted that the ratio estimator does not require that $N_{..k}$ or their sum $N_{..}$ be known. The Horvitz-Thompson estimator only requires that $N_{..}$ be known, whereas the post-stratified and linear regression estimators require that $\{N_{..k} | k = 1, \dots, K\}$ be known. In practice, the linear regression estimator has the additional complication that the covariance matrices \sum_{12} and \sum_{22} are unknown and must be estimated from the sample. In implementing \hat{Y}_{LR} in section 3, the known finite population quantities $(1/N_{..})N$ will be used in place of the limiting vector \mathbf{M}_2 .

2.2 Conditional Expectations and Variances of the Estimators

Using the asymptotic setup given earlier, the expectations and variances of the four estimators can be computed conditional on \hat{N} . For the case of post-stratification, conditioning on \hat{N} in a complex design is a natural extension of conditioning on the achieved post-stratum sample sizes in a simple random sample. In other situations, however, the question of what to condition on is a difficult one that may not have a unique answer (e.g., see Kiefer 1977). First, define the following three matrices:

$$\mathbf{H} = \sum_{12} \sum_{22}^{-1},$$

$$\mathbf{R} = \mathbf{H} - \mathbf{D}(\mu), \quad \text{and}$$

$$\mathbf{P} = \mathbf{H} - \mathbf{D}(\mu_k),$$

where $\mathbf{D}(\mu) = \text{diag}(\mu, \dots, \mu)$ and $\mathbf{D}(\mu_k) = \text{diag}(\mu_1, \dots, \mu_k)$ are $K \times K$ diagonal matrices. Below, we state the mean and variance of the four estimators without providing any details of the calculations. When the sample of first-stage units is large, each of the estimators has essentially the same conditional variance. The Horvitz-Thompson, ratio, and post-stratified estimators are, however, conditionally biased, whereas the linear regression estimator is not. Thus, the linear regression estimator has the smallest asymptotic mean square error among the four estimators considered here. Rao (1992) also noted the optimality of the regression estimator within a certain class of difference estimators and its negligible large sample bias.

(1) Horvitz-Thompson estimator:

$$E[\hat{Y}_{HT} | \hat{N}] = \mu + [\mathbf{1}' \mathbf{H} (\hat{N} - \mathbf{M}_2)]$$

$$\begin{aligned} \text{var}[\hat{Y}_{HT} | \hat{N}] &= m^{-1} [\mathbf{1}' (\sum_{11} - \sum_{12} \sum_{22}^{-1} \sum_{21}) \mathbf{1}] \\ &= m^{-1} [\mathbf{1}' \mathbf{V}_c \mathbf{1}] = V_{HT(c)}. \end{aligned}$$

(2) Ratio estimator:

$$\begin{aligned} E[\hat{Y}_R | \hat{N}] &= \mu + \left(\frac{N_{..}}{\hat{N}_{..}} \right) [\mathbf{1}' \mathbf{R} (\hat{N} - \mathbf{M}_2)] \\ &= \mu + [\mathbf{1}' \mathbf{R} (\hat{N} - \mathbf{M}_2)] + o(m^{-1}) \end{aligned}$$

$$\begin{aligned} \text{var}[\hat{Y}_R | \hat{N}] &= (N_{..}/\hat{N}_{..})^2 V_{HT(c)} \\ &= V_{HT(c)} + o(m^{-(3/2)}). \end{aligned}$$

(3) Post-stratified estimator:

$$\begin{aligned}
E[\hat{Y}_{PS} | \hat{N}] &= \mu + [r'P(\hat{N} - M_2)] \\
&= \mu + [1'P(\hat{N} - M_2)] + o(m^{-1}) \\
\text{var}[\hat{Y}_{PS} | \hat{N}] &= m^{-1}[r'V_c r] \\
&= V_{HT(c)} + o(m^{-(3/2)}).
\end{aligned}$$

(4) Linear regression estimator:

$$\begin{aligned}
E[\hat{Y}_{LR} | \hat{N}] &= \mu \\
\text{var}[\hat{Y}_{LR} | \hat{N}] &= V_{HT(c)}.
\end{aligned}$$

As noted in section 1, some minor modifications of the above formulas are necessary for designs, such as simple random sampling, in which $1'\hat{N} = N$. The derivation of the requisite modifications is straightforward and is not detailed here.

The large-sample biases of the first three estimators depend on $\hat{N} - M_2$. In other words, their biases are determined by how well the sample estimates the population distribution among the post-strata. In some special cases each of the first three can be conditionally unbiased. The post-stratified estimator, for example, will be approximately unbiased if $1'(H - D(\mu_k)) = 0'$. This occurs in simple random sampling and is possible, though certainly not generally true, in more complex designs. The matrix H can be interpreted as the slope in a multivariate regression of \hat{Y} on \hat{N} or of \bar{Y} on \bar{N} when the sample estimates are close to the population values. Thinking heuristically in superpopulation terms, if $E_\xi(Y_{ik}) = \mu_k N_{ik}$, as in Valliant (1993), with E_ξ denoting an expectation with respect to the model, then $E_\xi(Y_{.k}) = \mu_k N_{.k}$. The slope of the regression of $Y_{.k}$ on $N_{.k}$ is then μ_k and, in the unusual case in which the $\hat{Y}_{.k}$'s are independent, H is diagonal. In fact $H = D(\mu_k)$, so the conditional design-bias of the post-stratified estimator would be zero. If, on the other hand, the model has an intercept, *i.e.* if $E_\xi(Y_{.k}) = \alpha_k + \mu_k N_{.k}$, then the post-stratified estimator may have a substantial conditional design-bias. We will use this line of reasoning in the empirical study in section 3 to devise a population for which \hat{Y}_{PS} is conditionally biased.

Similar model-based thinking can be applied to the Horvitz-Thompson and ratio estimators to identify populations where the conditional design-biases will be predictably small for large samples. Suppose, as above, that the $\hat{Y}_{.k}$'s are independent. If each post-stratum total is unrelated to the number of units in the post-stratum, *i.e.* a peculiar situation in which $E_\xi(Y_{.k})$ does not depend on $N_{.k}$, then \hat{Y}_{HT} is conditionally design-unbiased. If $E_\xi(Y_{.k}) = \mu N_{.k}$, implying that all elementary population units have the same mean regardless of post-stratum, then \hat{Y}_R is conditionally design-unbiased.

2.3 Unconditional Expectations and Variances of the Estimators

Unconditionally, all estimators are approximately design-unbiased as noted below. The relative sizes of the variances depend on the values of Σ_{12} , Σ_{22} , μ , and $D(\mu_k)$. This is similar to the case of simple random sampling of a target y and an auxiliary x . In that case, whether the ratio estimator, $\bar{y}_s \bar{X}/\bar{x}_s$, or the regression estimator, $\bar{y}_s + b(\bar{X} - \bar{x}_s)$, has smaller design-variance also depends on the values of certain population parameters.

(1) Horvitz-Thompson estimator:

$$\begin{aligned}
E[\hat{Y}_{HT}] &= \mu \\
\text{var}[\hat{Y}_{HT}] &= m^{-1}[1'\Sigma_{11}1].
\end{aligned}$$

(2) Ratio estimator:

$$\begin{aligned}
E[\hat{Y}_R] &= \mu + o(m^{-1}) \\
\text{var}[\hat{Y}_R] &= m^{-1}[1'[\Sigma_{11} - 2\mu\Sigma_{21} + \mu^2\Sigma_{22}]1] \\
&\quad + o(m^{-(3/2)}).
\end{aligned}$$

(3) Post-stratified estimator:

$$\begin{aligned}
E[\hat{Y}_{PS}] &= \mu + o(m^{-1}) \\
\text{var}[\hat{Y}_{PS}] &= m^{-1}[1'[\Sigma_{11} - 2D(\mu_k)\Sigma_{21} \\
&\quad + D(\mu_k)\Sigma_{22}D(\mu_k)]1] + o(m^{-(3/2)}).
\end{aligned}$$

(4) Linear regression estimator:

The unconditional expectation and variance are the same as the conditional expectation and variance.

3. SIMULATION RESULTS

The theory developed in the preceding sections was tested in a set of simulation studies using three separate populations. The population size and basic sample design parameters for the three studies are listed in Table 1. The first population consists of a subset of the persons included in the first quarter sample of the 1985 National Health Interview Survey (NHIS) and the second population consists of a subset of the persons included in the September 1988 sample from the Current Population Survey (CPS). Both the NHIS and CPS are sample surveys conducted by the U.S. government. The variable of interest for the NHIS population is the number of restricted activity days in the two weeks prior to the interview and the variable of interest for the CPS population is weekly wages per person.

Table 1

Population Size and Basic Sample Design Parameters
for Three Simulation Studies

Population	Pop. Size N	No. of FSUs M	No. of sample FSUs m
HIS	2,934	1,100	115
CPS	10,841	2,826	200
Artificial	22,001	2,000	200

Post-strata in the NHIS and CPS populations were formed on the basis of demographic characteristics (as is typically done in household surveys) in order to create population sub-groups that were homogenous with respect to the variable of interest. For the NHIS population the variables age and sex were used to define 4 post-strata and for the CPS population the variables age, race, and sex were used to define 8 post-strata.

The third population is artificial; it was created with the intention of producing a substantial conditional bias in the post-stratified estimator of the mean. As noted in section 2.2, \hat{Y}_{PS} will be conditionally biased if the FSU post-stratum totals for the variable of interest, conditional on the number of units in each FSU/post-stratum, follow a model with a non zero intercept. With this in mind, we generated the population in such a way that

$$E(y_{ik} | N_{ik}) = \alpha_k + \beta N_{ik} + \gamma N_{ik}^2, \quad (4)$$

where N_{ik} is the number of units in the k^{th} post-stratum for the i^{th} FSU and α_k , β and γ are constants. Specifically, five post-strata were used with $\alpha_k = 100k$ ($k = 1, \dots, 5$), $\beta = 10$ and $\gamma = -.05$. In total two thousand FSUs were generated with the total number of units in the i^{th} FSU, say $N_{i\cdot}$, being a Poisson random variable with mean 10. Then, conditional on $N_{i\cdot}$, the numbers of units in the five post-strata (*i.e.*, $N_{i1}, N_{i2}, \dots, N_{i5}$) for the i^{th} FSU were determined using a multinomial distribution with parameters $N_{i\cdot}$ and $p_k = .20$ for $k = 1, 2, \dots, 5$.

For FSUs having $N_{ik} \geq 1$, the value of the variable of interest for the j^{th} unit in the k^{th} post-stratum for the i^{th} FSU was a realization of the random variable

$$y_{ijk} = \alpha_k/N_{ik} + \beta + \gamma N_{ik} + \epsilon_{1i} + \epsilon_{2ik} + \epsilon_{3ijk}N_{i\cdot} \\ (j = 1, \dots, N_{ik}; N_{ik} \geq 1),$$

where ϵ_{1i} , ϵ_{2ik} and ϵ_{3ijk} are three independent standardized chi-square (6 d.f.) random variables. This structure implies that $E(y_{ik} | N_{ik})$ is given by (4). Furthermore, the values of the variable of interest for units within an FSU

are correlated and the correlation depends upon whether the units are in the same post-stratum or not. This same algorithm was used in each of the 100 design strata. Twenty FSUs were generated in each design stratum giving a total of 2,000 FSUs.

A single-stage stratified design was used for the NHIS population with "households" being the FSUs. Ten design strata were used and an approximate 10% simple random sample of households was selected without replacement from each stratum. Each sample consisted of 115 households and each sample household was enumerated completely. A total of 5,000 such samples was selected for the simulation study.

Two-stage stratified sample designs were used for both the CPS and artificial populations. For the CPS population, geographic segments, employed in the original survey and composed of about four neighboring households, were used as FSUs and persons were the second-stage units. In both populations, 100 design strata were created with each stratum having approximately the same number of FSUs and a sample of $m = 2$ FSUs was selected with probability proportional to size from each stratum using the systematic sampling method described by Hansen, Hurwitz and Madow (1953, p. 343). Thus, 200 FSUs were selected for both populations. Second stage selection was also similar for both populations. For the CPS population a simple random sample of 4 persons was selected without replacement in each sample FSU having $N_{i\cdot} > 4$ and all persons were selected in each sample FSU where $N_{i\cdot} \leq 4$. For the artificial population the within FSU sample size was set at 15 rather than 4 which resulted in the complete enumeration of most sample FSUs. A total of 5,000 samples were selected from each of the populations for the simulation study.

In each sample, we computed \hat{Y}_{HT} , \hat{Y}_R , \hat{Y}_{PS} and two versions of \hat{Y}_{LR} . For the first version of the regression estimator, denoted $\hat{Y}_{LR}(\text{emp})$ in the tables, H was estimated separately from each sample as would be required in practice. Each component of Σ_{12} and Σ_{22} was estimated using the ultimate cluster estimator of covariance, appropriate to the design, as defined in Hansen, *et al.* (1953, p.419). The second version, denoted $\hat{Y}_{LR}(\text{theo})$, used the same value of H in each sample, which was an estimate more nearly equal to the theoretical value of the H matrix. For the CPS and artificial populations, the theoretical H matrix was estimated from empirical covariances derived from separate simulation runs of 5,000 samples. For the NHIS population the design was simple enough that a direct theoretical calculation of H was done. As the sample of FSUs becomes large, the performance of $\hat{Y}_{LR}(\text{emp})$ should approach that of $\hat{Y}_{LR}(\text{theo})$. The performance of $\hat{Y}_{LR}(\text{theo})$ is, consequently, a gauge of the best that can be expected from the empirical version of the regression estimator for a given sample size.

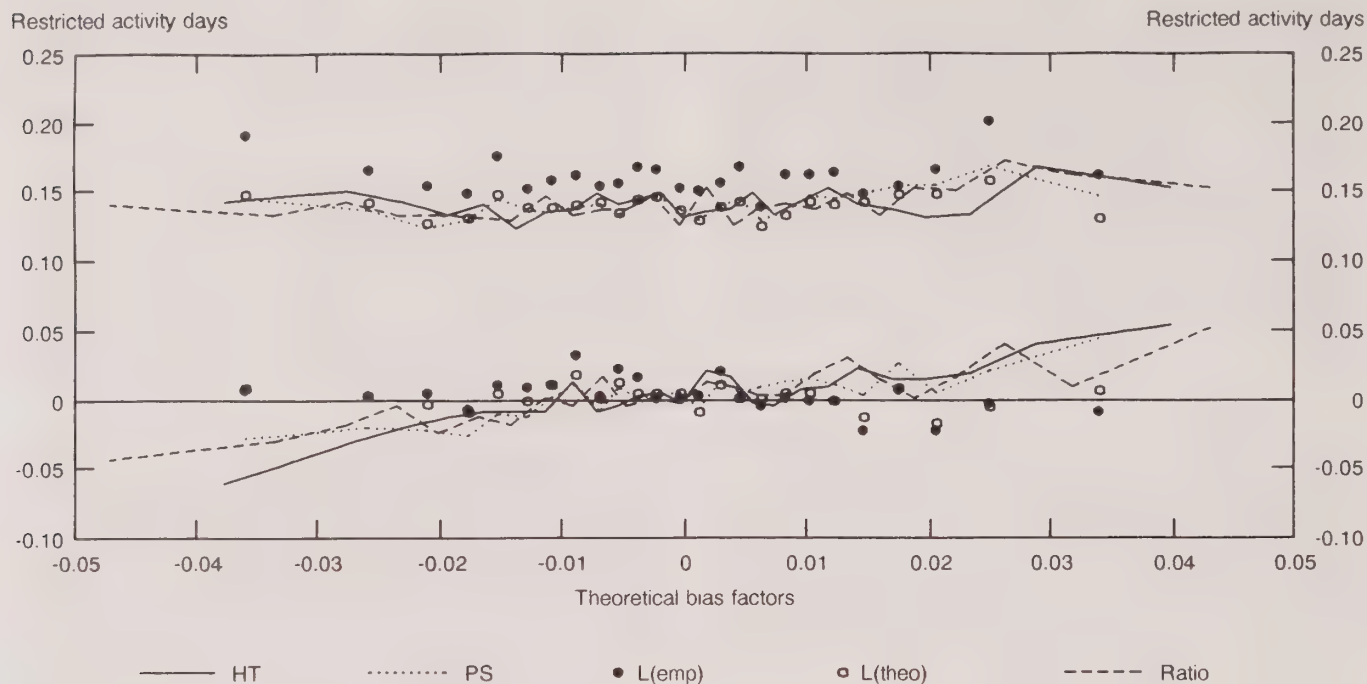


Figure 1. HIS simulation, $m = 115$

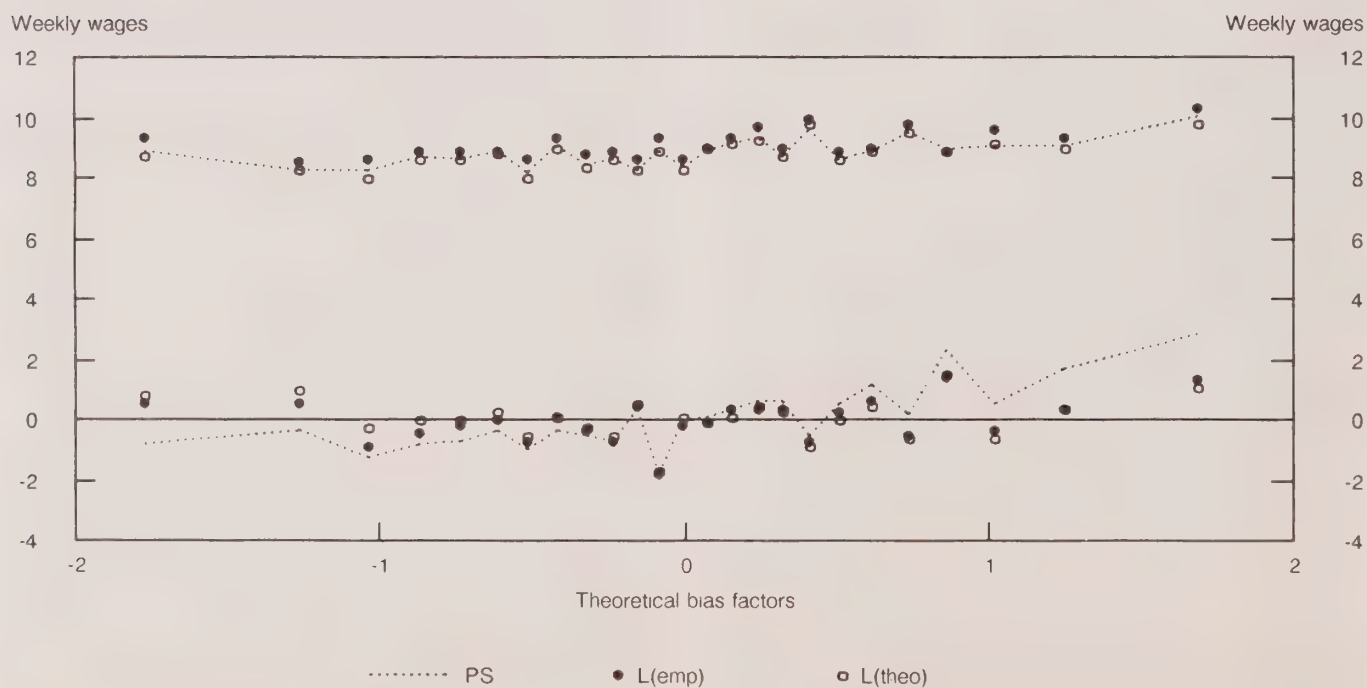


Figure 2. CPS simulation, $m = 200$

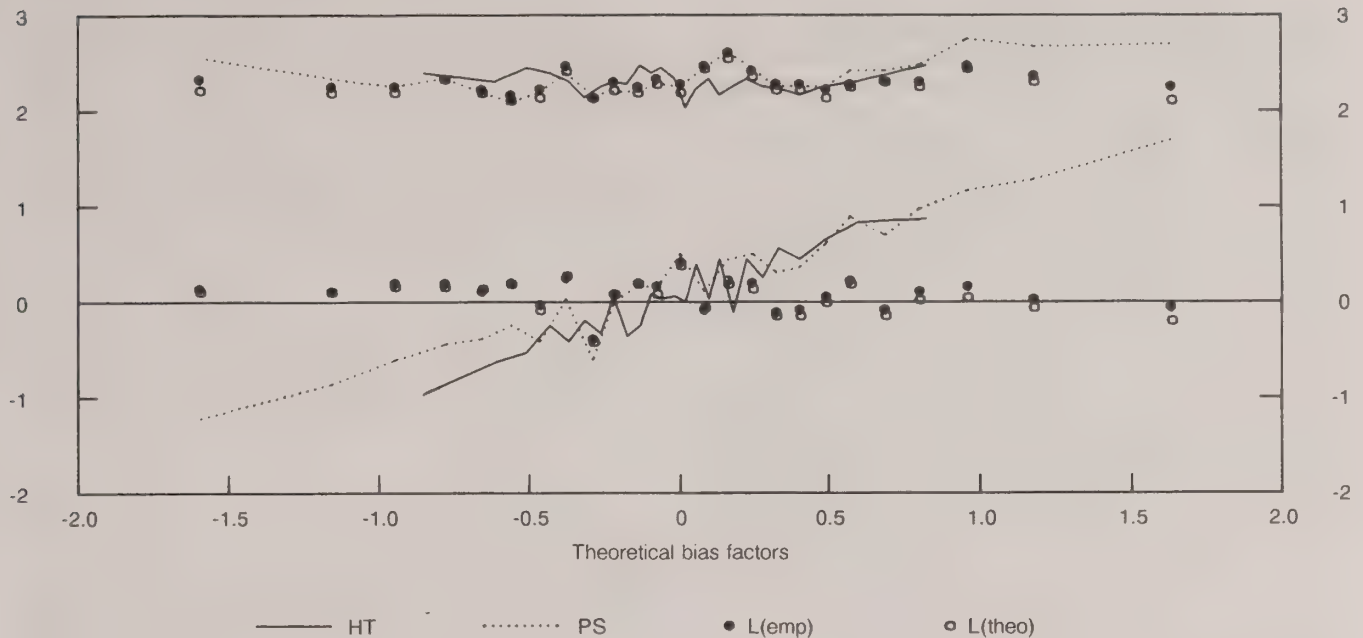


Figure 3. Artificial population simulation, $m = 200$

Table 2 lists unconditional results summarized over all 5,000 samples from each population. Empirical root mean square errors (rmse's) were calculated as $\text{rmse}(\hat{Y}) = [\sum_{s=1}^S (\hat{Y}_s - \bar{Y})^2 / S]^{1/2}$ with $S = 5,000$ and \hat{Y}_s being one of the estimates of the population mean from sample s . In the CPS and artificial populations, results for the Horvitz-Thompson and the ratio estimators were nearly identical so that only the former is shown. Across all samples, the bias of each of the estimators was negligible. As anticipated by the theory, $\hat{Y}_{LR}(\text{theo})$ was the most precise of the choices, although the largest gain compared to \hat{Y}_{PS} was only 4.7% in the artificial population. The need to estimate H destabilizes the regression estimator as shown in the results for $\hat{Y}_{LR}(\text{emp})$. For the NHIS and CPS populations, $\hat{Y}_{LR}(\text{emp})$ has a larger root mse than both $\hat{Y}_{LR}(\text{theo})$ and \hat{Y}_{PS} . The most noticeable loss is for the NHIS population where the root mse of $\hat{Y}_{LR}(\text{emp})$ is about 15% larger than that of either $\hat{Y}_{LR}(\text{theo})$ or \hat{Y}_{PS} . This result is consistent with the smaller FSU sample size and hence less stable estimate of H for the NHIS population.

Figures 1-3 present conditional simulation results. The 5,000 samples were sorted by the theoretical bias factors presented in section 2.2. The sorting was done separately for each of the estimators of the population mean. In the

cases of the two regression estimators, which are theoretically unbiased in large samples, the bias factor for \hat{Y}_{PS} was used for sorting. The sorted samples were then put into 25 groups of 200 samples each and empirical biases and root mse's were computed within each group. The group results were then plotted versus theoretical bias factors in the figures. The upper sets of points in each figure are the empirical root mse's of the groups, while the lower sets are empirical biases. The two regression estimators are conditionally unbiased as expected. The other estimators, however, have substantial conditional biases that, in the most extreme sets of samples, are important parts of the mse's. For the CPS population, the range of the bias factors for \hat{Y}_{HT} is so much larger (-10 to 10) than that of the other estimators that we have omitted \hat{Y}_{HT} from the plot for clarity. In the neighborhood of the balance point, $\hat{N} = \bar{N}$, all estimators perform about the same, but, because of a lack of data at the design stage, we have no control on how close to balance a particular sample may be. The safest choice for controlling conditional bias is, thus, $\hat{Y}_{LR}(\text{emp})$. This finding is similar to that of Valliant (1990), who noted that, in one-stage, stratified random or systematic sampling, the separate linear regression estimator is a good choice for controlling bias, conditional on the sample mean of an auxiliary variable.

Table 2

Simulation Results for Three Populations.
5,000 Samples were Selected from Each Population

Estimator	Rel-bias \hat{Y} (%)	rmse (\hat{Y})	$100^* \left[\frac{\text{rmse}(\hat{Y})}{\text{rmse}(\hat{Y}_{PS})} - 1 \right]$
HIS population			
\hat{Y}_{HT}	.12	.141	.05
\hat{Y}_R	.10	.141	.02
\hat{Y}_{PS}	.11	.141	0
$\hat{Y}_{LR}(\text{emp})$.19	.162	14.71
$\hat{Y}_{LR}(\text{theo})$.08	.140	-.96
CPS population			
\hat{Y}_{HT}	-.01	10.25	15.8
\hat{Y}_{PS}	0	8.85	0
$\hat{Y}_{LR}(\text{emp})$	-.03	9.11	3.0
$\hat{Y}_{LR}(\text{theo})$	-.01	8.79	-.6
Artificial population			
\hat{Y}_{HT}	.02	2.30	-2.93
\hat{Y}_{PS}	.12	2.37	0
$\hat{Y}_{LR}(\text{emp})$.04	2.31	-2.41
$\hat{Y}_{LR}(\text{theo})$.02	2.26	-4.70

4. DEFECTIVE FRAMES

The conditional biases discussed in the previous sections were of a technical, mathematical nature. A more serious, practical problem in many surveys, that can also lead to bias, is poor coverage of the target population; we address this situation in this section.

4.1 The Basic Problem of Defective Frames

In most real world applications not all of the elementary units in the population are included in the sampling frame. In household surveys, it is not unusual for some demographic subgroups, especially minorities, to be poorly covered by the sampling frame. Bailer (1989), for example, notes that in 1985 the sample estimate from the CPS of the total number of Black males, ages 22-24, was only 73% of an independent estimate of the total population of that group. Corresponding percentages for Black males, ages 25-29 and 60-61, were 80% and 76%.

To formalize the discussion of this type of coverage problem, suppose that N_k now refers to the number of elementary units in the frame and that \dot{N}_k is the *actual* number of population elements in the k^{th} post-stratum. In the discussion below terms with a dot on the top are population values while terms with no dot are frame values. Letting \dot{Y}_k be the aggregate of the y values over all population elements in the k^{th} post-stratum, then it follows that the *true population mean* is given by

$$\dot{\mu} = \lim_{L \rightarrow \infty} \frac{\sum_{k=1}^K \dot{Y}_k}{\sum_{k=1}^K \dot{N}_k} = \lim_{L \rightarrow \infty} \sum_{k=1}^K \frac{\dot{N}_k}{\dot{N}_\cdot} \frac{\dot{Y}_k}{\dot{N}_k} = \sum_{k=1}^K \dot{\phi}_k \dot{\mu}_k.$$

Obviously, all four of the estimators of the mean given in section 2 are biased (both conditionally and unconditionally) for $\dot{\mu}$; the additional bias term being given by $\mu - \dot{\mu}$ for all of the estimators. It should be noted that this bias term is $o(1)$ so it will dominate the other bias terms listed in section 2.2 as the number of FSUs increases. There is another even more basic problem; namely, in most cases the individual frame values N_k are not known so only the ratio estimator is well defined. For example, the Horvitz-Thompson estimator of the mean as defined in section 2 requires N_\cdot , the total number of units in the frame, but N_\cdot may be unknown. On the other hand, the \dot{N}_k (or least the proportions $\dot{\phi}_k$) may be known from independent sources and hence be available for the purposes of estimator construction. In household surveys, for instance, the \dot{N}_k may come from intercensal projections of population counts.

Before attempting to construct unbiased estimators for $\dot{\mu}$ it should be noted that

$$\begin{aligned} \mu - \dot{\mu} &= \sum_{k=1}^K (\phi_k - \dot{\phi}_k) (\mu_k - \dot{\mu}_k) \\ &\quad + \sum_{k=1}^K (\phi_k - \dot{\phi}_k) \dot{\mu}_k + \sum_{k=1}^K \dot{\phi}_k (\mu_k - \dot{\mu}_k). \end{aligned}$$

So, if we assume that for each post-strata the mean of the units in the frame is equal to the true population mean, (i.e. $\mu_k = \dot{\mu}_k$ for every k) then the bias term reduces to

$$\mu - \dot{\mu} = \sum_{k=1}^K (\phi_k - \dot{\phi}_k) \mu_k = \sum_{k=1}^K (\phi_k - \dot{\phi}_k) \dot{\mu}_k.$$

This is very strong (and also very expedient) assumption; however, addressing the problem of defective frame bias without such a condition is virtually impossible.

4.2 Alternative Estimators

The basic strategy is to construct an estimator for the defective frame bias, $\mu - \hat{\mu}$, and then subtract this estimator from the estimators studied earlier. Two cases need to be considered:

Case 1. The frame parameters $\{\phi_k, 1 \leq k \leq K\}$ are unknown, and

Case 2. The frame parameters $\{\phi_k, 1 \leq k \leq K\}$ are known.

Case 1. For this case only the ratio estimator is well defined and the only obvious candidate for an estimator of the bias is

$$\hat{B}_1 = \sum_{k=1}^K \left(\frac{\hat{N}_{\cdot k}}{\hat{N}_{\cdot \cdot}} - \phi_k \right) \frac{\hat{Y}_{\cdot k}}{\hat{N}_{\cdot k}} = \hat{Y}_R - \sum_{k=1}^K \phi_k \frac{\hat{Y}_{\cdot k}}{\hat{N}_{\cdot k}}.$$

Using the strategy given above, the resulting estimator for $\hat{\mu}$ is

$$\hat{Y}_1 = \hat{Y}_R - \hat{B}_1 = \sum_{k=1}^K \phi_k \frac{\hat{Y}_{\cdot k}}{\hat{N}_{\cdot k}}.$$

This is the “post-stratified” estimator usually found in practice. It is straightforward to verify the following properties of \hat{Y}_1 :

$$E[\hat{Y}_1 | \hat{N}] = \hat{\mu} + [p' P(\hat{N} - M_1)] + o(m^{-1})$$

where

$$p' = \begin{bmatrix} \phi_1 & \phi_2 & \dots & \phi_K \\ \phi_1 & \phi_2 & \dots & \phi_K \end{bmatrix}$$

$$\text{var}[\hat{Y}_1 | \hat{N}] = m^{-1} [p' V_c p] + o(m^{-(3/2)})$$

$$E[\hat{Y}_1] = \hat{\mu} + o(m^{-1})$$

$$\begin{aligned} \text{var}[\hat{Y}_1] &= m^{-1} [p' [\Sigma_{11} - 2D(\mu_k) \Sigma_{21} \\ &\quad + D(\mu_k) \Sigma_{22} D(\mu_k)] p] + o(m^{-(3/2)}). \end{aligned}$$

The attempt to correct for the defective frame bias is successful in the sense that \hat{Y}_1 is unconditionally unbiased for $\hat{\mu}$. However, the conditional bias is still present.

Case 2. For this case it can be verified that the estimator

$$\hat{B}_2 = (\mathbf{1} - p)' \left[\hat{Y} - \Sigma_{12} \Sigma_{22}^{-1} \left(\frac{\hat{N}}{\hat{N}_{\cdot \cdot}} - M_2 \right) \right],$$

is approximately, conditionally unbiased for $\mu - \hat{\mu}$ and, as \hat{Y}_{LR} is conditionally unbiased for μ , it follows directly that the estimator

$$\hat{Y}_2 = \hat{Y}_{LR} - \hat{B}_2 = p' \left[\hat{Y} - \Sigma_{12} \Sigma_{22}^{-1} \left(\frac{\hat{N}}{\hat{N}_{\cdot \cdot}} - M_2 \right) \right]$$

is both conditionally and unconditionally, approximately unbiased for $\hat{\mu}$. It can also be verified that

$$\text{var}[\hat{Y}_2 | \hat{N}] = \text{var}[\hat{Y}_2] = m^{-1} [p' V_c p].$$

In addition to the problems of the linear regression estimator cited earlier, this estimator is usually not even well defined as the frame parameters $\{\phi_k, 1 \leq k \leq K\}$ are rarely, if ever, known when the frame is defective.

5. CONCLUSION

This study has generalized the asymptotic techniques suggested by Robinson (1987) to study the problem of post-stratification from a design-based, conditional point-of-view. An important paper in the conditional study of post-stratification was that of Holt and Smith (1979), one of whose basic premises was that \hat{Y}_{PS} is conditionally unbiased. This will be true (at least asymptotically) only if $\mathbf{1}'(H - D(\mu_k)) = \mathbf{0}'$; so, in general, this premise is false. In fact, simple random sampling of elementary units may be one of the few realistic cases where this basic premise is true.

From a conditional point of view the linear regression estimator is preferable among the four studied here. Only the regression estimator is conditionally unbiased. The post-stratified estimator is no better (or worse) than either the Horvitz-Thompson or the ratio estimator; all have conditional bias terms of order $m^{-(1/2)}$. All of the estimators have the same conditional variance to terms of order m^{-1} ; furthermore, the conditional variance *does not* depend on \hat{N} , the vector of estimated proportions in the post-strata. Consequently, because of its conditional unbiasedness, the regression estimator has the smallest conditional mean square error.

The Horvitz-Thompson, ratio, and post-stratified estimators are unconditionally unbiased. Although somewhat illogical, one might attempt to make a case for the estimators by comparing their unconditional properties with the conditional properties of the linear regression estimator. But even from this mixed perspective, the $\hat{Y}_{LR}(\text{theo})$ estimator is clearly superior to the others. Not only is it conditionally unbiased, but the conditional variance of the linear regression estimator can be no larger than the unconditional variance of any of the other estimators. In large FSU samples, the empirical version of the regression estimator will inherit these good properties of $\hat{Y}_{LR}(\text{theo})$ and also perform well.

The problem of a defective frame introduces complications not found otherwise. Each of the estimators of the mean studied here is biased both conditionally and unconditionally. Bias adjustments are possible only under the restrictive assumption that the mean of units within each post-stratum is the same for all population units whether they are included or excluded from the frame.

An area we have not addressed is variance estimation. A design-based variance estimator for the regression estimator can be obtained using the methods of Särndal, Swensson and Wretman (1989).

ACKNOWLEDGMENT

Any opinions expressed are those of the authors and do not reflect policy of the U.S. Bureau of Labor Statistics. Also, the authors would like to thank the Associate Editor and the referee their germane and constructive comments; we feel they greatly strengthened our paper.

REFERENCES

- BAILAR, B. (1989). Information needs, surveys, and measurement errors. In *Panel Surveys*, (Eds. D. Kasprzyk, G. Duncan, G. Kalton and M.P. Singh). New York: Wiley.
- DURBIN, J. (1969). Inferential aspects of randomness of sample size in survey sampling. In *New Developments in Survey Sampling*, (Eds. N.L. Johnson and H. Smith). New York: Wiley.
- ERICSON, W.A. (1969). Subjective Bayesian models in sampling finite populations. *Journal of the Royal Statistical Society B*, 31, 195-233.
- FULLER, W.A. (1981). Comment on an empirical study of the ratio estimator and estimators of its variance by R.M. Royall and W.G. Cumberland. *Journal of the American Statistical Association*, 76, 78-80.
- HANSEN, M.H., HURWITZ, W.N., and MADOW, W.G. (1953). *Sample Survey Methods and Theory*, Vol. 1. New York: John Wiley and Sons.
- HANSEN, M.H., MADOW, W.G., and TEPPING, B.J. (1983). An evaluation of model-dependent and probability-sampling inferences in sample surveys. *Journal of the American Statistical Association*, 78, 776-796.
- HIDIROGLOU, M., and SÄRNDAL, C.-E. (1989). Small domain estimation: a conditional analysis. *Journal of the American Statistical Association*, 84, 266-275.
- HOLT, D., and SMITH, T.M.F. (1979). Post stratification. *Journal of the Royal Statistical Society A*, 142, 33-46.
- KIEFER, J. (1977). Conditional confidence statements and confidence estimators (with discussion). *Journal of the American Statistical Association*, 72, 789-827.
- KREWSKI, D., and RAO, J.N.K. (1981). Inference from stratified samples: Properties of the linearization, jackknife, and balanced repeated replication methods. *Annals of Statistics*, 9, 1010-1019.
- LITTLE, R.J.A. (1991). Post-Stratification: A modeler's perspective. *Proceeding of the Section on Survey Research Methods, American Statistical Association*, in press.
- RAO, J.N.K. (1985). Conditional inference in survey sampling. *Survey Methodology*, 11, 15-31.
- RAO, J.N.K. (1992). Estimating Totals and Distribution Functions Using Auxiliary Information at the Estimation Stage. Presented at the Workshop on Uses of Auxiliary Information in Surveys, Statistics Sweden.
- RAO, J.N.K., and WU, C.F.J. (1985). Inference from stratified samples: Second order analysis of three methods for nonlinear statistics. *Journal of the American Statistical Association*, 80, 620-630.
- ROBINSON, J. (1987). Conditioning ratio estimates under simple random sampling. *Journal of the American Statistical Association*, 82, 826-831.
- ROYALL, R.M. (1971). Linear regression models in finite population sampling theory. In *Foundations of Statistical Inference*, (Eds. V.P. Godambe and D.A. Sprott). Toronto: Holt, Rinehart, and Winston.
- SÄRNDAL, C.-E., SWENSSON, B., and WRETMAN, J. (1989). The weighted residual technique for estimating the variance of the finite population total. *Biometrika*, 76, 527-537.
- SÄRNDAL, C.-E., SWENSSON, B., and WRETMAN, J. (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.
- VALLIANT, R. (1990). Comparisons of variance estimators in stratified random and systematic sampling. *Journal of Official Statistics*, 6, 115-131.
- VALLIANT, R. (1993). Post-stratification and conditional variance estimation. *Journal of the American Statistical Association*, 88, 89-96.
- YATES, F. (1960). *Sampling Methods for Censuses and Surveys*, (3rd. Ed.). London: Griffin.

Sampling from Imperfect Frames with Unknown Amount of Duplication

SHIBDAS BANDYOPADHYAY and A.K. ADHIKARI¹

ABSTRACT

This study covers such imperfect frames in which no population unit has been excluded from the frame but an unspecified number of population units may have been included in the list an unspecified number of times each with a separate identification. When the availability of auxiliary information on any unit in the imperfect frame is not assumed, it is established that for estimation of a population ratio or a mean, the mean square errors of estimators based on the imperfect frame are less than those based on the perfect frame for simple random sampling when the sampling fractions of perfect and imperfect frames are the same. For estimation of a population total, however, this is not always true. Also, there are situations in which estimators of a ratio, a mean or a total based on smaller sampling fraction from imperfect frame can have smaller mean square error than those based on a larger sampling fraction from the perfect frame.

KEY WORDS: Imperfect frame; Efficiency.

1. INTRODUCTION

A frequent problem that arises while planning surveys is the non-availability of complete frames. The International Statistical Institute recognized the importance of studying the problem of sampling from imperfect frames and arranged discussions by experts on this topic during its 34th Session held in Ottawa, Canada where Hansen *et al.* (1963) and Szameitat and Schaffer (1963) presented invited papers. One may also refer to Singh (1977, 1983). Wright and Tsao (1983) have written a bibliography on frames to bring attention to problems which arise when sampling from imperfect frames.

Recently two separate surveys were undertaken by the Indian Statistical Institute to evaluate the impact of government sponsored programmes for the uplift of economic conditions of fishermen's community in West Bengal, India. In the first survey (1988), the households were selected using the membership registers of the Fishermen's Co-operative Societies (FCS). In the second and more recent survey, the list of beneficiary fishermen of the Fish Farmer's Development Agency (FFDA) was used. It was known that not all FCS members or FFDA beneficiaries would be from different households, but it was not possible to identify the FCS members or the FFDA beneficiaries belonging to the same household without contacting the households. Thus, when FCS membership registers or FFDA beneficiary lists were used for household selection, the frames contained an unknown number of duplication. Since the household information was collected by personal interview, it was possible to identify the duplication in the selected households only. The values of the

variables associated with the households in the sample were divided by the respective number of duplications in the frame while retaining the duplicate households in the sample under separate identification.

The set-up of imperfect frames discussed here is a special case of Rao (1968). One of the referees has pointed out that the situation discussed in the paper also occurs at Statistics Canada in certain frames for business surveys.

Imperfect frames to be covered in this study are those in which no population unit has been excluded from the frame but any population unit may have been included in the frame an unspecified number of times with a separate identification each time. It is assumed that it would be possible to ascertain, at the data collection stage, the number of duplicates in the frame for each selected unit. The possibility of selecting two or more duplicates of a population unit in the sample is not excluded. The availability of auxiliary information on the units in the imperfect frame is not assumed and only simple random sampling without replacement (SRSWOR) schemes are discussed.

Since the total number of population units will not be known from the imperfect frames to be covered here, problems of estimation of a mean of a population character and its total are not identical.

Here is the main question discussed in this paper. Which is better: to up-date the imperfect frame and select a sample, or to use the imperfect frame?

In the two surveys on fishermen's households, it was felt that most of the economic variables of interest would be highly related to the number of FCS members/FFDA beneficiaries in a household in the sense that the variability

¹ Shibdas Bandyopadhyay and A.K. Adhikari, Indian Statistical Institute, Calcutta, India 700 035.

of such an economic variable per FCS member/FFDA beneficiary would be less than the variability of the economic variable per household. It was felt that one could effectively use an imperfect frame in such situations.

It will be established that for situations such as above estimators of a ratio, a mean, or a total based on smaller sampling fraction, imperfect frame can have smaller Mean Square Error (MSE) than those based on a larger sampling fraction from the perfect frame.

Even when the variability is not related to the number of duplications as discussed above, it will be established that for estimating a ratio or a mean, using an imperfect frame will be preferable to using a perfect frame, from the MSE point of view, when the sampling fractions of the imperfect and the perfect frames are same.

2. NOTATIONS AND RELATIONS

Consider a finite population consisting of N units U_1, U_2, \dots, U_N . Let $U_1^*, U_2^*, \dots, U_M^*$ be the units listed in an imperfect frame. For $k = 1, 2, \dots, r$, let A_k denote the sub-population of the original N units consisting of N_k distinct population units. Each of the units in A_k is listed in the imperfect frame exactly k number of times under separate identifications. Assume that

- each U_i belongs to an A_k for some k , (i.e., each U_i is included in the imperfect frame at least once) and
- if U_j^* is selected in the sample using the imperfect frame, it will be possible to identify, at the data collection stage, the corresponding U_i and the associated value of k (i.e., the number of duplicates of U_i in the incomplete frame under separate identifications, one of which is the selected unit U_j^*) for which U_i belongs to A_k .

The following relations are valid.

$$N_1 + N_2 + \dots + N_r = N;$$

$$N_k \geq 0, k = 1, 2, \dots, r,$$

$$N_1 + 2N_2 + \dots + rN_r = M,$$

where r, N_1, N_2, \dots, N_r , and N are all unknown and only M is known with $M \geq N$; M may be written as, for unknown α ,

$$M = N(1 + \alpha), \quad \alpha \geq 0. \quad (2.1)$$

Let X and Y values on the unit U_i be X_i and Y_i respectively, ($i = 1, 2, \dots, N$). Since each U_j^* , ($j = 1, 2, \dots, M$), can be identified with a U_i for some i , ($i = 1, 2, \dots, N$), and since U_i belongs to A_k for some k , ($k = 1, 2, \dots, r$), define X, Y and C values for the unit U_j^* as

$$X_j^* = X_i/k, \quad Y_j^* = Y_i/k, \quad C_j^* = 1/k.$$

Because of assumptions (a) and (b), X^*, Y^* , and C^* values are observable for the selected units from the imperfect frame.

The following relations connect the measurements in the imperfect frame to those in the perfect frame.

$$\sum_{j=1}^M Y_j^* = M\bar{Y}^* = \sum_{i=1}^N Y_i = N\bar{Y};$$

$$\sum_{j=1}^M C_j^* = M\bar{C}^* = N;$$

$$\sum_{j=1}^M (Y_j^* - \bar{Y}^*)^2 = N\sigma_Y^2 - S(2, Y) + (N\bar{Y})^2(1/N - 1/M),$$

where

$$N\sigma_Z^2 = \sum_{i=1}^N (Z_i - \bar{Z})^2$$

and

$$S(a, Z) = \sum_{k=2}^r (1 - 1/k) \left\{ \sum_{i: U_i \in A_k} Z_i^a \right\}; \quad (2.2)$$

$$\sum_{j=1}^M (C_j^* - \bar{C}^*)^2 = N(1 - N/M) - S(0, Y);$$

$$\sum_{j=1}^M (Y_j^* - \bar{Y}^*)(C_j^* - \bar{C}^*) = N\bar{Y}(1 - N/M) - S(1, Y).$$

For the unit U_i let

$$D_i = Y_i - \bar{Y}; \quad W_i = Y_i - RX_i, \quad \text{where } R = \bar{Y}/\bar{X}. \quad (2.3)$$

Since no auxiliary information on the units is assumed, comparisons will be done on the basis of a SRSWOR sample. Let m be the size of the sample from the imperfect frame and n be the corresponding sample size had the frame been perfect. Define efficiency of a perfect frame compared to the corresponding imperfect frame, for any estimator, as

$$\rho = \frac{\text{MSE based on a sample of size } m \text{ from the imperfect frame}}{\text{MSE based on a sample of size } n \text{ had the frame been perfect}}. \quad (2.4)$$

Also define f as the common sampling fraction when the sampling fractions are same, *i.e.*,

$$n = fN, \quad m = fM = n(1 + \alpha). \quad (2.5)$$

3. RESULTS

Before we proceed to answer the main question raised in Section 1 on the choice of sampling from the perfect frame against sampling from the imperfect frame, we briefly look at the alternatives from cost considerations. If the total cost of up-dating the imperfect frame is expected to be more than the additional cost of data collection from the $(m - n)$ extra units, it is economical to use the imperfect frame with a larger sample size than to up-date the imperfect frame; this is so when

$$\frac{b_1}{b_0} \left(\frac{m - n}{N} \right) \leq 1, \quad (3.1)$$

where b_1 is the per-unit data collection cost and b_0 is the per-unit up-dating cost. It may be noted that one needs to visit effectively N units to up-date the incomplete frame since the remaining $(M - N)$ units are duplicates and can be identified because of assumption (b). It may also be noted that, even from a SRSWOR sample from the imperfect frame, the extra number of units to be canvassed is at most $(m - n)$ since the sample may contain the same unit under separate identifications. These observations lead to (3.1) for preference of using an imperfect frame.

As has been pointed out in Section 1, the total number of population units N will not be known from the imperfect frame. Thus the problems of estimation of a mean and a total are not identical; the problem of estimation of a mean essentially is the problem of estimation of a ratio, but a total can be estimated directly and unbiasedly, based on a SRSWOR sample of size m from the imperfect frame. It is thus appropriate to estimate a population ratio (similar to domain estimation) with estimation of a mean as a special case, and then to treat estimation of a total separately.

3.1 Estimation of a Ratio

For estimation of a ratio $R = (\bar{Y}/\bar{X})$, the usual ratio estimator is

$$\hat{R} = \bar{y}^*/\bar{x}^*,$$

where the lower case letters represent the corresponding quantities based on a sample, \bar{y}^* is the mean of Y^* values based on a sample of size m from the imperfect frame *etc.* \bar{y}^* and \bar{x}^* are respectively unbiased estimators of $(N\bar{Y}/M)$ and $(N\bar{X}/M)$. Using the delta method the MSE of \hat{R} , $E(\hat{R} - R)^2$, is given approximately by

$$\frac{M - m}{m(\bar{X}^*)^2(M - 1)M} \sum_{i=1}^M W_i^{*2}; \quad (3.2)$$

using the relations of Section 2, (3.2) can be rewritten as

$$\text{MSE}(\hat{R}) = \frac{M(M - m)}{m(N\bar{X})^2(M - 1)} \{N\sigma_W^2 - S(2, W)\},$$

where W values are defined in (2.3) and the W^* values correspondingly obtained. It follows from (2.2) that $S(2, W) \geq 0$, and hence from (3.2) one has

$$0 \leq 1 - \frac{S(2, W)}{N\sigma_W^2} \leq 1. \quad (3.3)$$

It now follows from (2.4) that efficiency ρ is

$$\rho = \frac{nM(M - m)(N - 1)}{mN(N - n)(M - 1)} \left\{ 1 - \frac{S(2, W)}{N\sigma_W^2} \right\}. \quad (3.4)$$

When sampling fractions are equal, ρ can be written as

$$\rho = \frac{(1 + \alpha)(N - 1)}{(1 + \alpha)(N - 1) + \alpha} \left\{ 1 - \frac{S(2, W)}{N\sigma_W^2} \right\}. \quad (3.5)$$

It, therefore, follows from (3.3) that ρ given by (3.5) satisfies

$$0 \leq \rho \leq 1 \quad (3.6)$$

and thus it is advantageous to use imperfect frame for estimation of a ratio.

It may be noted that $S(2, W)$ is nondecreasing in α and for fixed α , $S(2, W)$ has a larger value when the units with larger W values are replicated in the imperfect frame. Since σ_W^2 is fixed for a given set of N W values, there may be situations in which ρ in (3.4) is less than 1 (as a matter of fact $S(2, W)$ is equal to $N\sigma_W^2$ when W values are all equal and equal to 0) and consequently, there will be situations when sampling from imperfect frame will be preferable even with smaller sampling fraction to sampling from complete frame.

3.2 Estimation of a Mean

As seen in section 3.1, \bar{y}^* is an unbiased estimator of $(N\bar{Y})/M$ where M is known but N is unknown. Thus it is necessary to estimate N to get an estimator for \bar{Y} . It may be noted that \bar{c}^* is an unbiased estimator of (N/M) , and thus

$$\hat{\bar{Y}} = \bar{y}^*/\bar{c}^*$$

is a natural ratio-type estimator of \bar{Y} . On replacing \bar{x}^* in Section 3.1 by \bar{c}^* , the MSE of \hat{Y} is given by

$$\text{MSE}(\hat{Y}) = \frac{M(M-m)}{mN^2(M-1)} \{N\sigma_D^2 - S(2, D)\},$$

where D values are defined in (2.3). Replacing W in Section 3.1 by D we may conclude that (3.6) holds and imperfect frame is better when (2.5) is true.

3.3 Estimation of a Total

To estimate a total, say $N\bar{Y}$, based on a SRSWOR sample of size m from the imperfect frame, the usual estimator is

$$(\widehat{N\bar{Y}}) = M\bar{y}^*,$$

which is unbiased for $N\bar{Y}$, with variance

$$\begin{aligned} \text{MSE}(M\bar{y}^*) &= \text{Var}(M\bar{y}^*) \\ &= \frac{M(M-m)}{m(M-1)} \\ &\quad \left\{ N\sigma_Y^2 - S(2, Y) + (N\bar{Y})^2 \left(\frac{1}{N} - \frac{1}{M} \right) \right\}. \end{aligned}$$

One may write ρ as

$$\begin{aligned} \rho &= \frac{nM(M-m)(N-1)}{mN(N-n)(M-1)} \\ &\quad \left\{ 1 - \frac{S(2, Y) - (N\bar{Y})^2(1/N - 1/M)}{N\sigma_Y^2} \right\}. \end{aligned}$$

It is clear from the expression of $\text{Var}(M\bar{y}^*)$ that

$$\left\{ S(2, Y) - (N\bar{Y})^2 \left(\frac{1}{N} - \frac{1}{M} \right) \right\} / N\sigma_Y^2, \quad (3.7)$$

is less than or equal to unity. However, α and Y values may be so chosen that expression in (3.7) is negative. In such a case, even when (2.5) is true, imperfect frame with larger sampling fraction is inefficient. However, if the scatter of Y^* values are more homogeneous compared to Y values, i.e., if

$$\sum_{i=1}^N (Y_i - \bar{Y})^2 \geq \sum_{j=1}^M (Y_j^* - \bar{Y}^*)^2, \quad (3.8)$$

then the expression in (3.7) is always nonnegative. Now, one can draw similar conclusions as in Section 3.1, for example, (3.6) is valid when (2.5) is true.

4. AN ILLUSTRATION

As pointed out earlier, in the fishermen's survey, ultimate sampling units of beneficiary-fishermen were selected from the list of beneficiaries available. Being a multidisciplinary survey, many characteristics of the sampling units were observed from each of the sampling unit which either related to the household or to the fishing/fishery enterprise to which the sampling unit belonged. Since only the number of beneficiaries (M) was known and the number of corresponding households/enterprises (N) was not known, it was not possible to see the effect of using the imperfect frame for this survey. However for illustration in this paper, we take the samples drawn from one geographical area (a block within an administrative district in the West Bengal State) as our population and see the effect of resampling from it. In this area, there are 27 beneficiaries (M) and 23 distinct enterprises (N), 19 of the enterprises have single ownership (N_1) and 4 are of joint-ownership type (N_2). Our characteristics of interest are the cost of renovation of water areas (Y) and the acreage of operated water areas (X).

The summary statistics of Y and X are as follows:

$$\sum Y_i = 58,815, \quad \sum X_i = 23.36,$$

$$R = \left(\sum Y_i \right) / \left(\sum X_i \right) = 2,517.77,$$

$$S(2, Y) = 212,201,800, \quad S(2, D) = 145,101,018,$$

$$S(2, W) = 104,505,327,$$

$$23\sigma_Y^2 = 442,702,791, \quad 23\sigma_X^2 = 13.6503 \quad \text{and}$$

$$23\sigma_W^2 = 394,790,716,$$

where W is defined in (2.3).

To find the effect of sampling from the list of 27 beneficiaries we find estimates of

R = Renovation cost per acre of water area,

\bar{X} = Average water area per enterprise in acre and

$N\bar{X}$ = Total acreage of water areas operated by all 23 enterprises.

The table below gives the efficiencies for different choices of m and n .

Efficiency of sampling from perfect frame compared to sampling from imperfect frame (ρ)

Sample sizes		Efficiency for estimators of		
n	m	R	\bar{X}	$N\bar{X}$
2	2	0.8695	0.6453	0.9508
4	4	0.8841	0.6561	0.9668
6	6	0.9022	0.6696	0.9866
8	8	0.9225	0.6866	1.0117
8	9	0.7791	0.5781	0.8519
10	10	0.9551	0.7088	1.0444
10	11	0.8172	0.6065	0.8937

It can be seen that in most cases sampling from imperfect frame are more efficient.

ACKNOWLEDGEMENT

Authors wish to thank an Associate Editor and the referees for their valuable suggestions towards improvement of this paper.

REFERENCES

- HANSEN, M.H., HURWITZ, W.N., and JABINE, T.N. (1963). The Use of imperfect lists for probability sampling at the U.S. Bureau of Census. *Bulletin of the International Statistical Institute*, 40, 497-517, (with discussions).
- INDIAN STATISTICAL INSTITUTE (1988). *A study of Fishermen in West Bengal: 1985-1986*.
- RAO, J.N.K. (1968). Some non-response sampling theory when the frame contains an unknown amount of duplication. *Journal of the American Statistical Association*, 63, 87-90.
- SINGH, R. (1977). A note on the use of incomplete multi-auxiliary information in sample surveys. *Australian Journal of Statistics*, 19, 105-107.
- SINGH, R. (1983). On the use of incomplete frames in sample surveys. *Biometrical Journal*, 25, 545-549.
- SZAMEITAT, K., and SCHAFFER, K.A. (1963). Imperfect frames in statistics and the consequences for their use in sampling. *Bulletin of the International Statistical Institute*, 40, 517-538, (with discussions).
- WRIGHT, T., and TSAO, H.J. (1983). *A frame on frames: An annotated bibliography. Statistical Methods and Improvement of Data Quality*, (Ed. T. Wright). New York: Academic Press, 25-72.

An Alternative View of Forest Sampling

FRANCIS A. ROESCH, JR., EDWIN J. GREEN and CHARLES T. SCOTT¹

ABSTRACT

A generalized concept is presented for all of the commonly used methods of forest sampling. The concept views the forest as a two-dimensional picture which is cut up into pieces like a jigsaw puzzle, with the pieces defined by the individual selection probabilities of the trees in the forest. This concept results in a finite number of independently selected sample units, in contrast to every other generalized conceptualization of forest sampling presented to date.

KEY WORDS: Forest sampling; PPS sampling.

1. INTRODUCTION

The sampling of forests is often accomplished as a two part process: first a random point is located in the forest and then a cluster of trees in the vicinity of the point is selected for the sample by some rule. The two most common rules are known as (circular, fixed-area) plot sampling and (horizontal) point sampling. In the former, all trees for which the center of the cross-section of the bole at 4.5 feet above the ground is within a constant horizontal distance (d) of the random point are included in the sample. In the latter, tree i is selected for the sample if this center is within a horizontal distance αr_i of the random point, where r_i is the radius of the cross-section and α is a constant, chosen appropriately to obtain a desired sampling intensity. Tree i would be selected with probability proportional to πd^2 in plot sampling (the probability is the same for all trees) and with probability proportional to πr_i^2 (basal area of tree i) in point sampling (larger trees have a higher probability of selection).

There has been much discussion in the forestry literature about what the sample unit actually is in the various methods of forest sampling. The tree is considered the sample unit from one point of view (*e.g.* Oderwald 1981), while from other points of view, the cluster of trees associated with the point (*e.g.* Palley and Horwitz 1961; Schreuder 1970), the circular plot (*e.g.* Cunia 1965), and the point (*e.g.* Husch 1955) are considered the sample units. These various viewpoints are supported by different statistical tools. For example, treating the tree as the sample unit requires the use of finite population sampling theory, while considering the point as sample unit requires the use of the somewhat more advanced theory of infinite population sampling. In addition, plot sampling has traditionally been presented from the viewpoint of the plot as the sample unit, whereas point sampling has usually been

presented from the viewpoints of the tree or the point as the sample unit. Therefore, these very common and quite similar sampling mechanisms artificially appear disparate.

We will show a conceptualization of the primary sample unit that is applicable to every type of forest sampling scheme which selects trees based on the location of a random point. We will also show that this conceptualization is simple and that it provides a finite number of mutually exclusive and independently selected sample units. This is in contrast to the view of the tree or the cluster of trees as the sample unit, because trees are not selected independently and clusters of trees are not mutually exclusive. It also differs from the views of the randomly placed point or the plot as the sample unit, because there are an infinite number of units in these cases. We will also suggest that this alternative conceptualization is often more appropriate.

2. THE JIGSAW PUZZLE VIEW

Suppose that there are N trees in the forest with labels $1, 2, \dots, N$. Associated with the N trees are values of interest $\tilde{y} = \{\tilde{y}_1, \tilde{y}_2, \dots, \tilde{y}_N\}$, K -circles $K = \{K_1, K_2, \dots, K_N\}$, and selection areas of sizes $\tilde{A} = \{\tilde{A}_1, \tilde{A}_2, \dots, \tilde{A}_N\}$. Grosenbaugh and Stover (1957) first defined the K -circle in the context of point sampling. For our purposes the K -circle of tree i , K_i , is an imaginary circle, centered at tree center, with radius d in plot sampling and radius αr_i in point sampling. The selection area for tree i , of size A_i (in acres), is the portion of tree i 's K -circle which is within the forest, and is the area from within which a random point will select the tree for the sample.

When discussing point sampling, Palley and Horwitz (1961) contend that "... the primary sampling unit is a cluster of trees associated with a locus of origin. The locus

¹ Francis A. Roesch, Jr., Mathematical Statistician, Institute for Quantitative Studies, Southern Forest Experiment Station, USDA Forest Service, 701 Loyola Avenue, New Orleans, LA 70113; Edwin J. Green, Professor of Forestry, Cook College-Rutgers University, P.O. Box 231, New Brunswick, NJ 08903; and Charles T. Scott, Project Leader, Forest Ecosystem Modeling Unit, Northeastern Forest Experiment Station, USDA Forest Service, 359 Main Road, Delaware, Ohio 43015.

of origin is a point in the case of point sampling ...". Actually the locus of origin is not a point because the cluster of trees is not selected only from that point but rather from an infinite set of points within a specific area.

We offer the alternative view of the sample units being the mutually exclusive sections of ground resulting from the overlapping selection areas of the individual trees in the forest.

The treatment of the ground broken up into primary sampling units is clearly shown in Figure 1, for example. The correspondence between the population, sampling frame and sample unit as given in say Cochran (1977, p. 6) is apparent: the population (*or the puzzle picture*) is divided up into mutually exclusive, exhaustive sample units (*the puzzle pieces*) which together comprise the sample frame. Each ground segment has a definite probability of selection and the total of these probabilities over all segments is 1. We will call this the jigsaw puzzle view.

Associated with each ground segment are attributes of interest, the measurement of which will result in identical values from any point in that segment of ground. The crux of the matter is that individual points are equivalent within any particular segment. The ground segments, of course, are selected with probability proportional to size. In the case of point sampling, the segment size is determined by the basal areas and spatial distribution of the trees and the constant α chosen. Once α is chosen, the sample frame at a particular point in time is fixed. In the case of plot sampling, the size of the segment is determined by d and the spatial distribution of the trees. Thus, regardless of the

method used to determine the sample trees (*e.g.*, plot sampling or point sampling), all schemes can be thought of as cutting the puzzle up in some way, selecting the pieces with probability proportional to their size, and then turning each piece over to read the attributes associated with it.

Returning to our proposition that this view is often more appropriate, we note that the purpose of most forest surveys is to describe the *forest*, not the individual trees. Our aggregations are usually made on a per acre or hectare basis, *i.e.* units of the forest land, not units of the tree. From the same place we may measure many other things besides the trees such as topographic and site characteristics. It is therefore usually more appropriate to view pieces of the forest as the sample units rather than individual trees in the forest.

Although we will be working mostly in the context of forest sampling in general, our discussion is easily applied to any specific type of forest sampling which relies on the selection of trees by some function of randomly placed points. The only difference is the definition of the ground segments, or how we dissect the picture into puzzle pieces. For example, in plot sampling the ground is divided into pieces defined by overlapping circles of equal size, while in point sampling the definition is by overlapping circles of sizes proportional to each corresponding tree's basal area.

To examine this further, suppose that we randomly drop a point on the surface of a forest and use any function to select sample trees. Suppose also that within our forest are three trees (1, 2, and 3) whose selection areas overlap. In Figure 1, trees 1, 2 and 3 are centered at their respective numbers with their selection areas shown as circles. Each lettered segment represents a different sample unit. If the point falls in segment *a*, the empty cluster is chosen, in segment *b*, the cluster containing only tree 1, in segment *d*, the cluster of all three trees, *etc.* Tree 1 would therefore be selected from segments *b*, *c*, *d* or *e*. This results in a situation somewhat analogous to that described in Kish (1965, sec. 11.2), if we were to consider the tree to be the primary sample unit, in which a list to be sampled from contains duplicate listings of the same unit. In this case, the list would be one of clusters of trees, in which most trees are associated with more than one cluster. The clusters are selected with probability proportional to the size of the ground segment. The standard technique of weighting duplicate elements of a list, discussed by Kish, considers rather the selection of primary units with equal probability.

The jigsaw puzzle view reduces the complexity of the sampling mechanism in one sense by first mapping the tree population into the ground segment population and thereby reducing the sample list from a list of clusters of trees in which trees belong to more than one cluster to a list of unique ground segments. Our claim below that

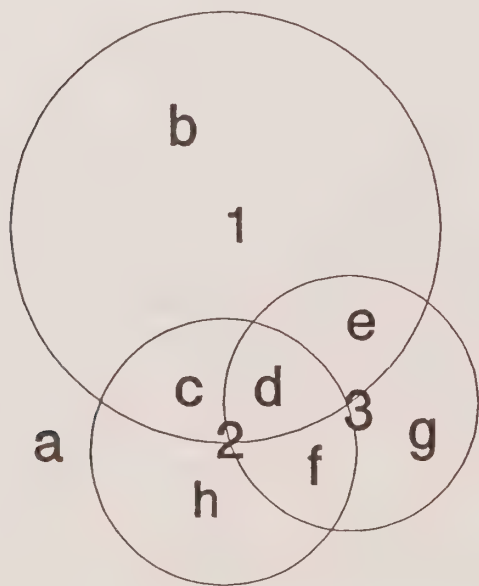


Figure 1. The Puzzle Pieces. Trees 1, 2 and 3 are centered at their respective numbers. The surrounding circles represent the selection areas of the trees. Each of the lettered segments represents a sample unit.

forest sampling simulations can be simplified by the jigsaw puzzle view is supported wholly by the tradeoff between the one time cost of this reduction in the complexity of the sample list and the need to select from that list many times.

To map the tree population into the segment population, an observation for a segment would preferably be the sum of weighted tree values, the weight for each tree being proportional to its probability of being observed from that particular segment. The probability that sampled tree i was selected from the particular ground segment j is:

$$p_{ij} = \left(\frac{A_j}{\tilde{A}_i} \right) Z_{ij},$$

where:

A_j = the area of segment j in acres, and

$Z_{ij} = \begin{cases} 1 & \text{if segment } j \text{ is part of the } k\text{-circle of tree } i \\ 0 & \text{otherwise.} \end{cases}$

The sum over j of p_{ij} is 1. We can now write the observation for each segment as a sum of weighted tree values:

$$y_j = \sum_{i=1}^N p_{ij} \tilde{y}_i. \quad (1)$$

Now suppose that we randomly drop m points on the surface of a forest with the same assumptions as above (our sampling is with replacement). An unbiased estimator of the total value of interest for a sample selected with probability proportional to size is:

$$\hat{Y} = \frac{A_T}{m} \sum_{j=1}^M \frac{y_j}{A_j} \quad (2)$$

$$= \frac{A_T}{m} \sum_{j=1}^M \frac{y_j}{A_j} W_j,$$

where:

$A_T = \sum_{j=1}^M A_j$; the total area of the forest in acres,

m = the number of sample points,

M = the number of ground segments, and

W_j = the number of times the j th unit appears in the sample.

Note that W_j is an integer between 0 and m , inclusive. A_j and y_j are fixed and W_j is random. In addition, we will define:

$Y = \sum_{i=1}^N \tilde{y}_i$; the total value of interest across all trees, and

$Y^* = \sum_{j=1}^M y_j$; the total value of interest across all segments.

To show that \hat{Y} is unbiased for Y , we will first show \hat{Y} to be unbiased for Y^* and then show that Y^* equals Y . Following Cochran (1977, p. 252-255), we can show \hat{Y} to be unbiased for Y^* :

$$E[\hat{Y}] = E \left[\frac{A_T}{m} \sum_{j=1}^M \frac{y_j}{A_j} W_j \right] \quad (3)$$

$$= \frac{A_T}{m} \sum_{j=1}^M \frac{y_j}{A_j} E[W_j].$$

W_j is a multinomial random variable and its expected value is equal to $m(A_j/A_T)$. Therefore

$$E[\hat{Y}] = \sum_{j=1}^M y_j = Y^*. \quad (4)$$

We can now show that \hat{Y} is unbiased for Y by showing that $Y^* = Y$. Substituting the right hand side of equation (1) for y_j in the definition of Y^* , we get:

$$Y^* = \sum_{j=1}^M \sum_{i=1}^N p_{ij} \tilde{y}_i. \quad (5)$$

After substituting in the definition of p_{ij} and rearranging the order of summation:

$$Y^* = \sum_{i=1}^N \tilde{y}_i \left[\frac{1}{\tilde{A}_i} \sum_{j=1}^M A_j Z_{ij} \right]. \quad (6)$$

Because

$$\tilde{A}_i = \sum_{j=1}^M A_j Z_{ij},$$

the term within the brackets on the right hand side of (6) equals 1, and

$$Y^* = \sum_{i=1}^N \tilde{y}_i = Y. \quad \text{Q.E.D.} \quad (7)$$

By definition, the variance of \hat{Y} is

$$V(\hat{Y}) = \left(\frac{1}{mA_T} \right) \sum_{j=1}^M A_j \left(\frac{A_T y_j}{A_j} - Y \right)^2. \quad (8)$$

The sample estimate of the variance is then (Cochran 1977):

$$v(\hat{Y}) = \frac{1}{m(m-1)} \sum_{j=1}^m \left(\frac{A_j y_j}{A_j} - \hat{Y} \right)^2. \quad (9)$$

The general development in equations (1) through (9) can be used for any specific type of forest sampling which follows the two part process of selecting trees from randomly placed points.

As a further example of the use of the jigsaw puzzle view, we will illustrate the sample frame when point samples are used to measure forest growth. For the greatest efficiency, measurements are taken at two points in time and the same random points are used both times. This type of sampling for forest growth is known as remeasured point sampling and has been discussed at length in the literature, most recently by Van Deusen *et al.* (1986) and Roesch *et al.* (1989, 1991, 1993). If a remeasured point sample had been taken, and Figure 1 represented time 1, the puzzle for the overall sample might be cut up into pieces like those in Figure 2. Trees 1, 2 and 3 are the same as those in Figure 1 and tree 4 is a tree which grew into the stand between times 1 and 2. The inner circles represent the trees' point sample areas of selection at time 1

(say αr_{i1} , including a subscript for time) and the outer circles represent the point sample areas of selection at time 2 (αr_{i2} is larger due to an increase in basal area). Tree 4 only has an outer circle since it did not exist at time 1 and tree 2 only has an inner circle since it died prior to time 2. The dotted circle represents the selection area tree 2 would have had at time 2 if time 2 had occurred just prior to the tree's demise. Therefore, the dotted circle does not contribute to the definition of the segments.

If the random point lands in segment *a*, trees 1 and 3 would be measured at both times and tree 2 would be measured only at time 1; in segment *b*, tree 1 would be measured at both times and tree 3 would only be measured at time 2. This exemplifies the fact that even though another dimension was added to the sample (the time dimension), the forest sample concept remains the same, since the time dimension can be collapsed down onto the puzzle picture. So, in addition to the conditions mentioned above, the definition of the segments depends upon the exact times of each measurement. This concept of the sample unit is helpful in understanding the estimators of the components of change from time 1 to time 2 given in Van Deusen *et al.* (1986) and Roesch *et al.* (1989 and 1991).

3. DISCUSSION

Given the simplicity of the jigsaw puzzle concept, one might wonder why this view of forest sampling has not been proposed before. The most compelling reason is probably that the above estimators cannot be calculated when the A_j 's are unknown. Since a particular tree's area of selection might be divided between many of the puzzle pieces and the size of a particular puzzle piece may be limited by trees not sampled by that piece, the selection areas of both sample and non-sample trees must be known to calculate the A_j 's of the selected segments. For example, referring to Figure 1, if our point landed in section *c*, we would sample trees 1 and 2 and the area of *c* + *d* would be readily calculable. However, to calculate \hat{Y} and $v(\hat{Y})$, we need the area of *c* alone, for which we do not have adequate sample information. We will show that this apparent deficiency is unimportant by showing that \hat{Y} can be reexpressed in terms which are calculable. This will, in fact, always be the case no matter which sampling method is described by the jigsaw puzzle view.

The jigsaw puzzle view of point sampling is actually a mapping of the tree population into the associated ground segment population. We can reexpress \hat{Y} to show that it is equivalent to the usual point sampling estimator which is based upon the tree population. Expanding equation (2) to include the definition of y_j and subsequent rearrangement gives:

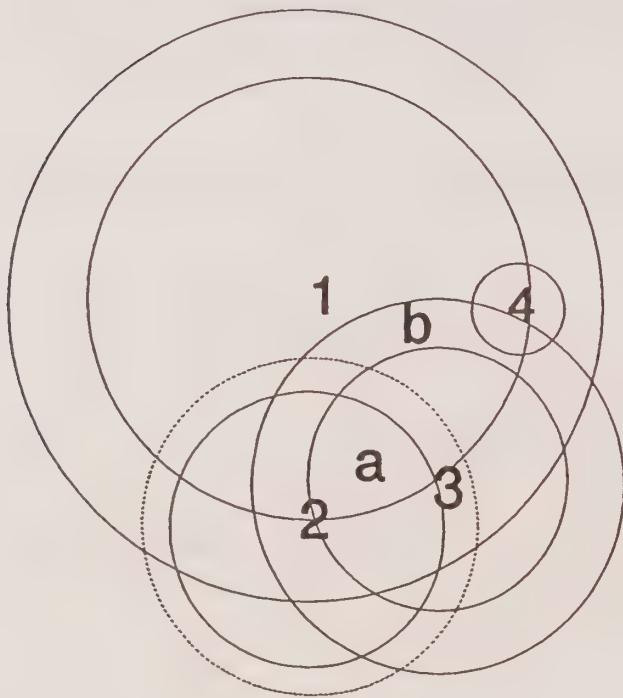


Figure 2. Puzzle pieces defined by location, size, and time. An example of sample units in a remeasured point sample. Trees 1 and 3 have grown and survived, tree 2 grew somewhat before dying and tree 4 is ingrowth.

$$\begin{aligned}
\hat{Y} &= \frac{A_T}{m} \sum_{j=1}^M \frac{y_j}{A_j} w_j \\
&= \frac{A_T}{m} \sum_{j=1}^M \frac{\sum_{i=1}^N p_{ij} \tilde{y}_i}{A_j} w_j \\
&= \frac{A_T}{m} \sum_{j=1}^M \sum_{i=1}^N \frac{\tilde{y}_i z_{ij} w_j}{\bar{A}_i} \quad (10) \\
&= \frac{A_T}{m} \sum_{i=1}^N \frac{\tilde{y}_i}{\bar{A}_i} \sum_{j=1}^M z_{ij} w_j \\
&= \frac{A_T}{m} \sum_{i=1}^N \frac{\tilde{y}_i}{\bar{A}_i} w_i,
\end{aligned}$$

where w_i equals the number of times tree i is selected for the sample. The final expression in (10) is the usual point sample estimator.

The purpose of this paper, therefore, is not to introduce a new set of estimators for sampling systems which already have reasonably good estimators, but rather to show how sampling schemes of quite disparate justifications in the literature are related in general. This alternative avenue of understanding may be useful in many ways. For one, we believe that some abstract forest sampling systems may be easier to understand if put into the framework described above. Our experience is that students, for instance, readily grasp the idea of point sampling when taught as merely a method of dividing the forest up into non-overlapping jigsaw puzzle pieces which are then sampled with probability proportional to size. Researchers who are interested in developing new forest sampling schemes or new estimators for existing schemes may benefit from this view because it provides another path for understanding new sampling schemes and for programming the forest sampling simulations used to test the new methods. The simulation discussed in Roesch (1993), for example, was simplified by using the jigsaw puzzle view rather than the other conceptualizations of the forest sampling frame which had been suggested up to that time. The simplification stemmed from the fact that the bulk of the simulation could be used for many different sampling schemes with only minor modifications to the subroutine which dissected the puzzle.

Because forest sampling simulations often start with a mapped forest, the A_j 's are readily obtainable. Once the puzzle is dissected, y_j can be calculated for each piece. The simulator then simply selects these pieces from a list in proportion to their size. Contrast this with the simulation resulting from the view of the point as the sample unit. In this latter simulation, a random point would be dropped and the tree list searched for all of the trees close enough to that point to be selected for the sample. Then the attributes of interest would be calculated. Since the probability of selecting a point from an infinite population twice is zero, this list search and calculation would have to be repeated for each random point, possibly resulting in repeated calculation of the attributes from the same cluster of trees. For simulation purposes, the optimal approach to programming will depend upon the length of the tree list to be searched, the degree of clustering in the tree population, and the number of random points.

4. CONCLUSION

We've presented a generalized forest sampling concept which utilizes a finite number of ground segments as the sample units existing within a land-area based sample frame. We have also given estimators based on this concept. The jigsaw puzzle view should be of help in understanding the similarities and differences between different methods of forest sampling by putting all of the methods into the same framework. Although we would not normally utilize the associated estimators in their given form in an actual forest survey, we can always find an equivalent calculable form. The additional benefit of an alternative route for sampling simulations is not only one of academics but also economics. Given the amount of time and money it takes to acquire data in forestry studies, the ability to easily test the properties of different sampling methods before they are applied in the field is of paramount importance. We would not endeavor to undermine the importance of a thorough theoretical development of proposed forest sampling schemes as the crucial first step, but simulation of these schemes before implementation may help uncover overlooked problems. This alternative conceptualization will, in general, facilitate comparisons within any group of forest sampling schemes.

ACKNOWLEDGEMENTS

The authors would like to thank the Associate Editor, two anonymous referees and Hans Schreuder for their helpful comments.

REFERENCES

- COCHRAN, W.G. (1977). *Sampling Techniques*, (3rd Ed.). New York: John Wiley.
- CUNIA, T. (1965). Continuous forest inventory, partial replacement of samples and multiple regression. *Forest Science*, 11, 480-502.
- GROSENBAUGH, L.R., and STOVER, W.S. (1957). Point-sampling compared with plot-sampling in southeast Texas. *Forest Science*, 3, 2-14.
- HUSCH, B. (1955). Results of an investigation of the variable plot method of cruising. *Journal of Forestry*, 53, 570-574.
- KISH, L. (1965). *Survey Sampling*. New York: John Wiley.
- ODERWALD, R.G. (1981). Point and plot sampling – the relationship. *Journal of Forestry*, 79, 377-378.
- PALLEY, M.N., and HORWITZ, L.G. (1961). Properties of some random and systematic point sampling estimators. *Forest Science*, 7, 52-65.
- ROESCH, F.A. Jr. (1993). Adaptive cluster sampling for forest inventories. *Forest Science*, 39. In press.
- ROESCH, F.A. Jr., GREEN, E.J., and SCOTT, C.T. (1989). New compatible estimators for survivor growth and ingrowth from remeasured horizontal point samples. *Forest Science*, 35, 281-293.
- ROESCH, F.A. Jr., GREEN, E.J., and SCOTT, C.T. (1991). Compatible basal area and number of trees estimators from remeasured horizontal point samples. *Forest Science*, 37, 136-145.
- ROESCH, F.A. Jr., GREEN, E.J., and SCOTT, C.T. (1993). A test of alternative estimators for volume at time 1 from remeasured point samples. *Canadian Journal of Forest Research*, 23, 598-604.
- SCHREUDER, H.T. (1970). Point sampling theory in the framework of equal-probability cluster sampling. *Forest Science*, 16, 240-246.
- VAN DEUSEN, P.C., DELL, T.R., and THOMAS, C.E. (1986). Volume growth estimation from permanent horizontal points. *Forest Science*, 32, 415-422.

Panel Surveys: Adding the Fourth Dimension

GRAHAM KALTON and CONSTANCE F. CITRO¹

ABSTRACT

Surveys across time can serve many objectives. The first half of the paper reviews the abilities of alternative survey designs across time – repeated surveys, panel surveys, rotating panel surveys and split panel surveys – to meet these objectives. The second half concentrates on panel surveys. It discusses the decisions that need to be made in designing a panel survey, the problems of wave nonresponse, time-in-sample bias and the seam effect, and some methods for the longitudinal analysis of panel survey data.

KEY WORDS: Panel surveys; Rotating panel surveys; Repeated surveys; Panel attrition; Time-in-sample bias; Seam effect; Longitudinal analysis.

1. INTRODUCTION

Survey populations are constantly changing over time, both in composition and in the characteristics of their members. Changes in composition occur when members enter the survey population through birth (or reaching adulthood), immigration, or leaving an institution (for a noninstitutional population) or leave through death, emigration, or entering an institution. Changes in characteristics include, for example, a change from married to divorced, or from a monthly income of \$2,000 to one of \$2,500. These population changes give rise to a range of objectives for the analysis of survey data across time. This paper reviews survey designs that produce the data needed to satisfy these various objectives.

The paper is divided into two parts. The first part contains a review of the general issues involved in conducting surveys across time, including the objectives of such surveys and the types of survey design that may be employed. This part is to be found in Section 2. The second, and main, part of the paper discusses one particular survey design, a panel survey that follows the same sample of units through time. The considerations involved in designing, conducting, and analyzing a panel survey are reviewed in Section 3. Section 4 provides some concluding remarks.

2. SURVEYS ACROSS TIME

This section presents an overview of analytic objectives across time, of designs for surveys across time, and of the extent to which different designs can satisfy the various objectives. The discussion relies heavily on Duncan and Kalton (1987), which contains a more detailed treatment of these issues.

Changes in population characteristics and composition over time lead to a variety of objectives for surveys across time. These objectives include the following:

- (a) The estimation of population parameters (*e.g.*, the proportion of the population in poverty) at distinct time points.
- (b) The estimation of average values of population parameters across time (*e.g.*, the daily intake of iron averaged across a year).
- (c) The estimation of net changes, that is changes at the aggregate level (*e.g.*, the change in the proportion of unemployed from one month to the next).
- (d) The estimation of gross changes and other components of individual change (*e.g.*, the proportion of persons who were in poverty one year and were not in poverty in the following year).
- (e) The aggregation of data for individuals over time (*e.g.*, the summation of twelve monthly incomes to give annual income).
- (f) The collection of data on events occurring in a specified time period (*e.g.*, becoming unemployed), and on their characteristics (*e.g.*, duration of spells of unemployment).
- (g) The cumulation of samples over time, especially samples of rare populations (*e.g.*, women who become widowed).
- (h) The maintenance of a sample of members of a rare population that was identified at one point of time (*e.g.*, scientists and engineers identified from a large-scale survey at one point of time).

¹ Graham Kalton, Westat, 1650 Research Blvd., Rockville, Maryland, U.S.A. 20850; Constance F. Citro, National Research Council, 2101 Constitution Ave. N.W., Washington, D.C., U.S.A., 20418.

A number of survey designs have been developed to provide the data needed to address these objectives. These designs are:

- *Repeated survey.* A repeated survey is a series of separate cross-sectional surveys conducted at different time points. No attempt is made to ensure that any of the same elements are sampled for the individual surveys. The elements are sampled from a population defined in the same manner for each individual survey (e.g., the same geographical boundaries and age-limits) and many of the same questions are asked in each survey.
- *Panel survey.* A panel survey collects the survey data for the same sample elements at different points of time.
- *Repeated panel survey.* A repeated panel survey is made up of a series of panel surveys each of a fixed duration. There may be no overlap in the time period covered by the individual panels, for instance one panel may start only as (or after) the previous one ends, or there may be an overlap, with two or more panels covering part of the same time period.
- *Rotating panel survey.* Strictly, a rotating panel survey is equivalent to a repeated panel survey with overlap. Both limit the length of a panel, and have two or more panels in the field at the same time. However, it seems useful to distinguish between the two designs because they have different objectives. Rotating panel surveys are widely used to provide a series of cross-sectional estimates and estimates of net change (e.g., of unemployment rates and changes in such rates), whereas repeated panel surveys with overlaps also have a major focus on longitudinal measures (e.g., durations of spells of unemployment). In consequence, repeated panel surveys tend to have longer durations and have fewer panels in operation at any given time than rotating panel surveys.
- *Overlapping survey.* Like a repeated survey, an overlapping survey is a series of cross-sectional surveys conducted at different time points. However, whereas the repeated survey does not attempt to secure any sample overlap from the survey at one time point to the next, an overlapping survey is designed to provide such overlap. The aim may be to maximize the degree of sample overlap while taking into account both the changes desired in selection probabilities for sample elements that remain in the survey population and also changes in population composition over time.
- *Split panel survey.* A split panel survey is a combination of a panel survey and a repeated survey or rotating panel survey.

The choice of design in a particular case depends on the objectives to be satisfied. Some designs are better than others for some objectives but poorer for other objectives. Some designs cannot satisfy certain objectives at all. For a detailed discussion, see Duncan and Kalton (1987).

The strength of a repeated survey is that it selects a new sample at each time point, so that each cross-sectional survey is based on a probability sample of the population existing at that time. A panel survey is based on a sample drawn from the population existing at the start of the panel. Although attempts are sometimes made to add samples of new entrants to a panel at later time points, such updating is generally difficult to do and is done imperfectly. Moreover, nonresponse losses from a panel as it ages heighten concerns about nonresponse bias when the panel sample is used to estimate cross-sectional parameters for later time points. For these reasons, repeated surveys are stronger than panel surveys for producing cross-sectional and average cross-sectional estimates (objectives (a) and (b)). With average cross-sectional estimates, another factor to be considered is the correlation between the values of the survey variables for the same individual at different time points. When this correlation is positive, as it generally is, it increases the standard errors of the average cross-sectional estimates from a panel survey. This factor thus also favours repeated surveys over panel surveys for average cross-sectional estimates.

The superior representation of the samples for a repeated survey at later time points also argues in favour of a repeated survey over a panel survey for estimating net change (assuming that the interest in net change relates to changes in both population composition and characteristics). However, in this case the positive correlations of the values of the survey variables for the same individuals across time decreases the standard errors of estimates of net change from a panel survey. Hence the presence of this correlation operates in favour of the panel design for measuring net change.

The key advantages of the panel design are its abilities to measure gross change, and also to aggregate data for individuals over time (objectives (d) and (e)). Repeated surveys are incapable of satisfying these objectives. The great analytic potential provided by the measurement of individual changes is the major reason for using a panel design.

Repeated surveys can collect data on events occurring in a specified period and on durations of events (e.g., spells of sickness) by retrospective questioning. However, retrospective questioning often introduces a serious problem of response error in recalling dates, and the risk of telescoping bias. A panel survey that uses a reference period for the event that corresponds to the interval between waves of data collection can eliminate the telescoping problem by using the previous interview to bound the recall (i.e., an illness reported at the current interview can be discarded if it had already been reported at the previous one). Similarly, a panel survey can determine the duration of an event from successive waves of data collection, limiting the length of recall to the interval between waves.

Repeated data collections over time can provide a vehicle for accumulating a sample of members of a rare population, such as persons with a rare chronic disease or persons who have recently experienced a bereavement. Repeated surveys can be used in this manner to generate a sample of any form of rare population. Panel surveys, however, can be used to accumulate only samples of new rare events (such as bereavements) not of stable rare characteristics (such as having a chronic disease). If a sample of members with a rare stable characteristic (e.g., persons with doctoral degrees) has already been identified, a panel survey can be useful for maintaining the sample over time, with suitable supplementation for new entrants at later waves (for an example, see Citro and Kalton 1989).

Rotating panel surveys are primarily concerned with estimating current levels and net change (objectives (a) and (c)). As such, elements are usually retained in the panel for only short periods. For instance, sample members remain in the monthly Canadian Labour Force Survey for only six months. The extent to which individual changes can be charted and aggregation over time can be performed is thus limited by the short panel duration. A special feature of rotating panel surveys is the potential to use composite estimation to improve the precision of both cross-sectional estimates and estimates of net change (see Binder and Hidioglou 1988; Cantwell and Ernst 1993). See also Fuller *et al.* (1993) for an alternative method of using past information in forming estimates from a rotating panel design.

Like rotating panel surveys, overlapping surveys are primarily concerned with estimating current levels and net change. They can also provide some limited information on gross change and aggregations over time. Overlapping survey designs are applicable in situations where some sample overlap is required and where the desired element selection probabilities vary over time. This situation arises in particular in establishment surveys, where the desired selection probability for an establishment may vary from one cross-sectional survey to the next to reflect its change in size and type of activity. In such circumstances, a Keyfitz-type procedure can be applied to maximize the retention of elements from the previous survey while taking account of changes in selection probabilities and population composition (see, for example, Keyfitz 1951; Kish and Scott 1971; Sunter 1986). The U.S. Internal Revenue Service Statistics of Income Division's corporate sample provides an example of an overlapping survey design (Hinkins *et al.* 1988).

By combining a panel survey with a repeated survey or a rotating panel survey, a split panel survey can provide the advantages of each. However, given a constraint on total resources, the sample size for each component is necessarily smaller than if only one component had been used. In particular, estimates of gross change and other measures of individual change from a split panel survey

will be based on a smaller sample than would have been the case if all the resources had been devoted to the panel component.

In comparing alternative designs for surveys across time, the costs of the designs need to be considered. For instance, panel surveys avoid the costs of repeated sample selections incurred with repeated surveys, but they face costs of tracking and tracing mobile sample members and sometimes costs of incentives to encourage panel members to continue to cooperate in the panel (see Section 3). If two designs can each satisfy the survey objectives, the relative costs for given levels of precision for the survey estimates need to be examined.

3. PANEL SURVEYS

The repeated measures over time on the same sampled elements that are obtained in panel surveys provide such surveys with a key analytic advantage over repeated surveys. The measurements of gross change and other components of individual change that are possible with panel survey data form the basis of a much greater understanding of social processes than can be obtained from a series of independent cross-sectional snapshots. The power of longitudinal data derived from panel surveys has long been recognized (see, for instance, Lazarsfeld and Fiske 1938; Lazarsfeld 1948), and panel surveys have been carried out in many fields for many years. Subjects of panel surveys have included, for example, human growth and development, juvenile delinquency, drug use, victimizations from crime, voting behaviour, marketing studies of consumer expenditures, education and career choices, retirement, health, and medical care expenditures. (See Wall and Williams (1970) for a review of early panel studies on human growth and development, Boruch and Pearson (1988) for descriptions of some U.S. panel surveys, and the Subcommittee on Federal Longitudinal Surveys (1986) for descriptions of U.S. federal panel surveys.) In recent years, there has been a major upsurge in interest in panel surveys in many subject-matter areas, and especially in household economics. The ongoing U.S. Panel Study of Income Dynamics began in 1968 (see Hill 1992 for a description of the PSID) and similar long-term panel studies have been started in the past decade in many European countries. The U.S. Bureau of the Census started to conduct the Survey of Income and Program Participation (SIPP) in 1983 (Nelson *et al.* 1985; Kasprzyk 1988; Jabine *et al.* 1990), and Statistics Canada introduced the Survey of Labour and Income Dynamics (SLID) in 1993. The growth in interest in panel surveys has also given rise to an increase in literature about the methodology of such surveys, including such recent texts as Kasprzyk *et al.* (1989), Magnusson and Bergman (1990), and Van de Pol (1989).

This section reviews the major issues involved in the design and analysis of panel surveys. The treatment is geared towards repeated panel surveys of fixed duration like the SIPP and SLID, but most of the discussion applies more generally to all forms of panel survey.

3.1 Design Decisions for a Panel Survey

The time dimension adds an extra dimension of complexity to a panel survey as compared with a cross-sectional survey. In addition to all the decisions that need to be made about the design features of a cross-sectional survey, a wide range of extra decisions needs to be reached for a panel survey. Major design decisions include:

- *Length of the panel.* The longer the panel lasts, the greater is the wealth of data obtained for longitudinal analysis. For instance, the longer the panel, the greater the number of spells of unemployment starting during the life of the panel that will be completed before the end of the panel, and hence the greater the precision in estimating the survival function for such spells. On the other hand, the longer the panel, the greater the problems of maintaining a representative cross-sectional sample at later waves, because of both sample attrition and difficulties in updating the sample for new entrants to the population.

It can sometimes be beneficial to vary the length of the panel between different types of panel members. Thus, for instance, when the analytic objectives call for it, panel members with certain characteristics (*e.g.*, members of a minority population) or who experience certain events during the course of the regular panel (*e.g.*, a divorce) can be retained in the panel for extended periods of observation.

- *Length of the reference period.* The frequency of data collection depends on the ability of respondents to recall the information collected in the survey over time. Thus, the PSID, with annual waves of data collection, requires recall of events occurring in the previous calendar year, whereas SIPP, with four-monthly waves of data collection, requires recall for the preceding four months. The longer the reference period, the greater the risk of recall error.
- *Number of waves.* In most cases the number of waves of data collection is determined by a combination of the length of the panel and the length of the reference period. The greater the number of waves, the greater the risk of panel attrition and time-in-sample effects, and the greater the degree of respondent burden.
- *Overlapping or non-overlapping panels.* With a repeated panel survey of fixed duration, a decision needs to be made as to whether the panels should overlap across time. Consider, for instance, the proposal of a National Research Council study panel that the SIPP should be a four-year panel (Citro and Kalton 1993). One possibility

is to run each panel for four years, starting a new panel when the previous one finishes. Another possibility is run each panel for four years, but starting a new panel every two years. Yet another possibility is to run each panel for four years, starting a new panel every year.

The design of nonoverlapping panels has the benefit of simplicity, since only one panel is in the field at any one time. It also produces a large sample for longitudinal analysis; for instance, the panels with the nonoverlapping design can be roughly twice the size of those with the design that has two overlapping panels at any one time. However, this increase in sample size for nonoverlapping panels does not apply for cross-sectional estimates, since the data from the panels covering a given time point can be combined for cross-sectional estimation. Also, the cross-sectional estimates for a time period near the end of a panel with the nonoverlapping design are at greater risk of bias from attrition, time-in-sample bias, and failure to update the sample fully for new population entrants than is the case with an overlapping design, in which one panel is of more recent origin. Moreover, the overlapping design permits the examination of such biases through a comparison of the results for the two panels for a given time period, whereas no such examination is possible with a nonoverlapping design. Another limitation of the nonoverlapping design is that it may not be well positioned to measure the effect of such events as a change in legislation. For instance, if legislation takes effect in the final year of a nonoverlapping panel, there will be little opportunity to evaluate its effect by comparing the situations of the same individuals before and for some period after the legislation is enacted. With overlapping panels, one of the panels will provide a wider window of observation.

- *Panel sample size.* For a given amount of annual resources, the sample size for each panel is determined by the preceding factors. A larger panel for longitudinal analysis can be achieved by lengthening the reference period and by employing a nonoverlapping design. The sample size for cross-sectional estimates can be increased by lengthening the reference period, but not by using a nonoverlapping design.

The above list determines the major parameters of a panel survey design, but there still remain a number of other factors that need to be considered:

- *Mode of data collection.* As with any survey, a decision needs to be made as to whether the survey data are to be collected by face-to-face interviewing, by telephone, or by self-completion questionnaire, and whether computer assisted interviewing (CAPI or CATI) is to be used. With a panel survey, this decision needs to be made for each wave of data collection, with the possibility of different modes for different waves (for instance, face-to-face

interviewing at the first wave to make contact and establish rapport, with telephone interviewing or mail questionnaires at some of the later waves). When modes may be changed between waves, consideration needs to be given to the comparability of the data across waves. Sometimes a change in mode may involve a change in interviewer, as for instance would occur with a change from face-to-face interviewing to a centralized CATI operation. Then the effects of a change of interviewer between waves on the respondent's willingness to continue in the panel and on the comparability of responses across waves also need to be carefully considered.

- *Dependent interviewing.* With panel surveys there is the possibility of feeding back to respondents their responses at earlier waves of data collection. This dependent interviewing procedure can secure more consistent responses across waves, but risks generating an undue level of consistency. The ease of application of dependent interviewing depends on the length of the interval between waves and the mode of data collection. Processing the responses from one wave to feed back in the next is easier to accomplish if the interval between waves is a long one and if computer assisted interviewing is employed. Edwards *et al.* (1993) describe the use of dependent interviewing with CAPI in the Medical Care Beneficiary Survey, a survey which involves three interviews per year with each respondent.
- *Incentives.* Monetary or other incentives (*e.g.*, coffee mugs, calculators, lunch bags) may be offered to sampled persons to encourage their participation in a survey. With a panel survey, incentives may be used not only to secure initial participation but also to maintain cooperation throughout the duration of the panel. There is an issue of when are the best times to provide incentives in a panel survey (*e.g.*, at the first wave, at an intermediate wave, or at the last wave of the panel). Panel survey researchers often send respondents a survey newsletter, frequently giving some recent highlights from the survey findings, at regular intervals, both to generate goodwill for the survey and to maintain contact with respondents (see below). Birthday cards sent at the time of the respondents' birthdays are also often used for these purposes.
- *Respondent rules.* Survey data are often collected from proxy informants when respondents are unavailable for interview. With a panel survey, this gives rise to the possibility that the data may be collected from different individuals at different waves, thus jeopardizing the comparability of the data across waves. The respondent rules for a panel survey need to take this factor into account.
- *Sample design.* The longitudinal nature of a panel survey needs to be considered in constructing the sample design for the initial wave. Clustered samples are commonly employed for cross-sectional surveys with face-to-face

interviewing in order to reduce fieldwork travel costs and to enable frame construction of housing unit listings to be performed only for selected segments. These benefits are bought at the price of the increase in the variance of survey estimates arising from the clustering. The optimum extent of clustering depends on the various cost factors involved and the homogeneity of the survey variables in the clusters (see, for instance, Kish 1965). With a panel survey, the use and extent of any clustering should be determined in relation to the overall panel with all its waves of data collection. In particular, the benefit of reduced fieldwork costs disappears for waves of data collection that are conducted by telephone interviewing or mail questionnaire. Also the migration of panel members to locations outside the original clusters reduces the benefit of the initial clustering for fieldwork costs at later waves. (However, some benefits of the initial clustering still operate for the large proportion of mobile persons who move within their own neighbourhoods.)

Oversampling of certain population subgroups is widely used in cross-sectional surveys to provide sufficient numbers of subgroup members for separate analysis. Such subgroups may, for instance, comprise persons with low incomes, minority populations, persons in a specified age-group, or persons living in certain geographical areas. Such oversampling can also be useful in panel surveys, but caution is needed in its application. With long-term panels, one reason for caution is that the objectives of the survey may change over time. Oversampling to meet an objective identified at the start of a panel may prove harmful to objectives that emerge later. Another reason for caution is that many of the subgroups of interest are transient in nature (*e.g.*, low income persons, persons living in a given geographical area). Oversampling persons in such subgroups at the outset of the panel may be of limited value for later waves: some of those oversampled will leave the subgroup while others not oversampled will join it. Thirdly, the definition of the desired subgroup for longitudinal analysis needs to be considered. For instance, SIPP data are used to estimate durations of spells on various welfare programs. Since such estimates are usually based on new spells starting during the life of the panel, it may not be useful to oversample persons already enrolled on welfare programs. See Citro and Kalton (1993) for a discussion of oversampling for the SIPP.

When oversampling of a certain subgroup of the population (*e.g.*, a minority population) is desired for a panel survey, the oversampling may require a large screening operation. The assessment of the cost of such screening should be made in the context of the full panel with all its waves of data collection. An expensive screening operation at the first wave may well be justifiable in this context.

- *Updating the sample.* When the sole objective of a panel survey is longitudinal analysis, it may be sufficient to adopt a cohort approach that simply follows the initial sample selected for the first wave. However, when cross-sectional estimates are also of interest, it may be necessary to update the sample at each wave to represent new entrants to the population. Updating for all types of new entrants is often difficult, but it is sometimes possible to develop fairly simple procedures to account for certain types of new entrants. For instance, in a panel of persons of all ages, babies born to women panel members after the start of the panel can be included as panel members. The SIPP population of inference comprises persons aged 15 and over. By identifying in initial sampled households persons who are under 15 years old but who will attain that age before the end of the panel, by following them during the panel, and by interviewing them after they reach 15 years of age, a SIPP panel can be updated for this class of new entrants (Kalton and Lepkowski 1985).

Attention also needs to be paid to panel members who leave the survey population. For some the departure is clearly permanent (*e.g.*, deaths), but for others it may be only temporary (*e.g.*, going abroad or entering an institution). If efforts are made to keep track of temporary leavers, they can be readmitted to the panel if they return to the survey's population of inference.

Panel surveys such as SIPP and PSID collect data not only for persons in original sampled households, but also for other persons – nonsampled persons – with whom they are living at later waves. The prime purpose of collecting survey data for nonsampled persons is to be able to describe the economic and social circumstances of sampled persons. The issue arises as to whether any or all nonsampled persons should remain in the panel after they stop living with sampled persons. For some kinds of analysis it is useful to follow them. However, to follow them would eat significantly into the survey's resources.

When data are collected for nonsample members, these data may be used simply to describe the circumstances of sample members, in which case analyses are restricted to sample members, with nonsample members being assigned weights of zero. Alternatively, nonsample members can be included in cross-sectional analyses. In this case appropriate weights for sample and nonsample persons need to be developed to reflect the multiple ways in which individuals may appear in the dataset. Huang (1984), Ernst (1989) and Lavallée and Hunter (1993) describe the fair share weighting approach that may be used for this purpose.

- *Tracking and tracing.* Most panel surveys encounter the problem that some panel members have moved since the last wave and cannot be located. There are two ways to try to handle this problem. First, attempts can be made

to avoid the problem by implementing procedures for tracking panel members between waves. One widely-used procedure when there is a long interval between waves is to send mailings, such as birthday cards and survey newsletters, to respondents between waves, requesting the post office to provide notification of change of address if applicable. Another tracking device is to ask respondents for the names and addresses or telephone numbers of persons close to them (*e.g.*, parents) who are unlikely to move and who will be able to provide locating information for them if they move.

The second way to deal with lost panel members is to institute various tracing methods to try to locate them. With effort and ingenuity, high success rates can be achieved. Some methods of tracing may be specific for the particular population of interest (*e.g.*, professional societies for persons with professional qualifications) while others may be more general, such as telephone directories, computerized telephone number look-ups, reverse telephone directories for telephone numbers of neighbours, mail forwarding, marriage licence registers, motor vehicle registrations, employers, and credit bureaus. It can be useful to search death records for lost panel members, particularly for long-term panel surveys. Panel members found to have died can then be correctly classified, rather than being viewed as non-respondents. Methods of tracing are discussed by Burgess (1989), Clarridge *et al.* (1978), Crider *et al.* (1971) and Eckland (1968).

3.2 Problems of Panel Surveys

Panel surveys share with all surveys a wide range of sources of nonsampling error. This section does not review all these sources, but rather concentrates on three sources that are unique to panel surveys, namely wave nonresponse, time-in-sample bias and the seam effect.

3.2.1 Wave nonresponse

The nonresponse experienced by panel surveys at the first wave of data collection corresponds to that experienced by cross-sectional surveys. The distinctive feature of panel surveys is that they encounter further nonresponse at subsequent waves. Some panel members who become non-respondents at a particular wave do not respond at any subsequent wave while others respond at some or all subsequent waves. The former are often termed attrition cases and the latter non-attrition cases. The overall wave nonresponse rates in panel surveys increase with later waves, but with well-managed surveys the rate of increase usually declines appreciably over time. For example, with the 1987 SIPP panel, the sample loss was 6.7% at wave 1, 12.6% at wave 2, and it then increased slowly to 19.0% at wave 7 (Jabine *et al.* 1990). The tendency for the nonresponse rate to flatten off at later waves is comforting,

but nevertheless the accumulation of nonresponse over many waves produces high nonresponse rates at later waves of a long-term panel. For instance, in 1988, after 21 annual rounds of data collection, the PSID nonresponse rate for individuals who lived in 1968 sampled households had risen to 43.9% (Hill 1992).

The choice between the two standard general-purpose methods for handling missing survey data – weighting adjustments and imputation – is not straightforward for wave nonresponse in panel surveys. For longitudinal analysis, the weighting approach drops all records with one or more missing waves from the data file and attempts to compensate for them by weighting adjustments applied to the remaining records. This approach can lead to the loss of a substantial amount of data when the data file covers several waves. On the other hand, the imputation approach retains all the reported data, but requires conducting wholesale imputations for missing waves. A compromise approach uses imputation for some patterns of wave nonresponse (e.g., those with only one missing wave, where data are available from both adjacent waves), and weighting for others (see, for example, Singh *et al.* 1990). For cross-sectional analysis, separate data files may be created for each wave. These files can comprise all the respondents for that wave, with either weighting adjustments or imputations for the wave nonrespondents. Kalton (1986) and Lepkowski (1989) discuss general methods for handling wave nonresponse, Lepkowski *et al.* (1993) discuss imputations for wave nonresponse in the SIPP, and Michaud and Hunter (1993) describe plans for handling wave nonresponse in the SLID.

With wave nonresponse there is the possibility of collecting some or all of the data for the missing wave at a subsequent interview. However, the quality of the retrospective data collected in this way needs to be carefully assessed. An experiment was conducted to examine the utility of this approach with the 1984 SIPP panel, using a missing wave form to collect responses for a skeleton set of core questions for the missing wave (Huggins 1987; Singh 1993). The analyses showed substantially fewer transitions in receipt of income, assets, and government assistance from the missing wave form than from benchmark data. In consequence the use of the missing wave form was discontinued. Administrative records may sometimes provide another possible source of skeletal data for missing waves.

3.2.2 Time-in-sample bias

Time-in-sample bias, or panel conditioning, refers to the effect that panel members' responses at a given wave of data collection are affected by their participation in previous waves. The effect may reflect simply a change in reporting behaviour. For example, a respondent may recognize from previous interviews that a "Yes" response

to a question leads to follow-up questions, whereas a "No" answer does not. The respondent may therefore give a "No" answer to avoid the burden of the extra questions. Alternatively, a respondent may learn from previous interviews that detailed information on income is needed, and may therefore prepare for later interviews by collecting the necessary data. The time-in-sample effect may also reflect a change in actual behaviour. For example, a respondent may enroll in the food stamp program as a result of learning of its existence from the questions asked about it at earlier waves of data collection.

A recent experimental study of panel conditioning in a four-year panel study of newlyweds found some evidence that participation in the study did affect marital well-being (Veroff *et al.* 1992). However, that study used in-depth interviewing techniques that are more intrusive than those used in most surveys. A number of studies of panel conditioning that have been conducted in more standard survey settings have found that conditioning effects do sometimes occur, but they are not pervasive (Traugott and Katosh 1979; Ferber 1964; Mooney 1962; Waterton and Lievesley 1989).

A benefit of rotating and overlapping panel surveys is that they enable estimates for the same time period obtained from different panels to be compared. Such comparisons have clearly identified the presence of what is termed "rotation group bias" in the U.S. and Canadian Labour Force Surveys (e.g. Bailer 1975, 1989, and U.S. Bureau of the Census 1978, for the U.S. Current Population Survey; Ghangurde 1982, for the Canadian Labour Force Survey). Rotation group bias may reflect nonresponse bias and conditioning effects. In analyses comparing the overlapping 1985, 1986 and 1987 SIPP panels, Pennell and Lepkowski (1992) found few differences in the results from the different panels.

3.2.3 Seam effect

Many panel surveys collect data for subperiods within the reference period from the last wave of data collection. The SIPP, for instance, collects data on a monthly basis within the four-month reference period between waves. The seam effect refers to the common finding with this form of data collection that the levels of reported changes between adjacent subperiods (e.g., going on or off of a welfare program from one month to the next) are much greater when the data for the pair of subperiods are collected in different waves than when they are collected in the same wave. The seam effect has been found to be pervasive in SIPP, and to relate to both reciprocity status and amounts received (see, for example, Jabine *et al.* 1990; Kalton and Miller 1991). It has also been found in PSID (Hill 1987). Murray *et al.* (1991) describe approaches used to reduce the seam effect in the Canadian Labour Market Activity Survey.

3.3 Longitudinal Analysis

There is a substantial and rapidly expanding literature on the analysis of longitudinal data, including a number of texts on the subject (*e.g.* Goldstein 1979; Hsiao 1986; Kessler and Greenberg 1981; Markus 1979). This treatment cannot be comprehensive, but rather identifies a few general themes.

- *Measurement of gross change.* As has already been noted, a key analytic advantage of a panel survey over a repeated survey is the ability to measure gross change, that is, change at the individual level. The basic approach to measuring gross change is the turnover table that tabulates responses at one wave against the responses to the same question at another wave. The severe limitation to this form of analysis is that changes in measurement errors across waves can lead to serious bias in the estimation of the gross change (for further discussion, see Kalton *et al.* 1989; Rodgers 1989; Abowd and Zellner 1985; Chua and Fuller 1987; Fuller 1990; and Skinner 1993).
- *Relationship between variables across time.* Panel surveys collect the data necessary to study the relationships between variables measured at different times. For instance, based on the data collected in the 1946 British birth cohort, the National Survey of Health and Development, Douglas (1975) found that children who were hospitalized for more than a week or who had repeated hospitalizations between the ages of 6 months and 3½ years exhibited more troublesome behaviour in school and lower reading scores at age 15. In principle, cross-section surveys may employ retrospective questions to collect the data needed to perform this type of analysis. However, the responses to such questions are often subject to serious memory error, and potentially to systematic distortions that affect the relationships investigated.
- *Regression with change scores.* Regression with change scores can be used to avoid a certain type of model misspecification. Suppose that the correct regression model for individual i at time t is

$$Y_{it} = \alpha + \beta x_{it} + \gamma z_{it} + \epsilon_{it},$$

where x_{it} is an explanatory variable that changes value over time and z_{it} is an explanatory variable that is constant over time (*e.g.*, gender, race). Suppose further that z_{it} is unobserved; it may well be unknown. Then β can still be estimated from the regression on the change scores:

$$Y_{i(t+1)} - Y_{it} = \beta(x_{i(t+1)} - x_{it}) + \epsilon_{i(t+1)} - \epsilon_{it},$$

(Rodgers 1989; Duncan and Kalton 1987).

- *Estimation of spell durations.* The data collected in panel surveys may be used to estimate the distribution of lengths of spells of such events as being on a welfare program. In panel surveys like the SIPP, some individuals have a spell in progress at the start of the panel (initial-censored spells), some start a spell during the panel, and some spells continue beyond the end of the panel (right-censored spells). Thus, not all spells are observed in their entirety. The distribution of spell durations may be estimated by applying survival analysis methods, such as the Kaplan-Meier product-limit estimation procedure to all new spells (including right-censored new spells) starting during the life of the panel (*e.g.* Ruggles and Williams 1989).
- *Structural equation models with measurement errors.* The sequence of data collection in a panel survey provides a clear ordering of the survey variables that fits well with the use of structural equation modelling for their analysis. This form of analysis can make allowance for measurement errors, and with several repeated measures can handle correlated error structures (*e.g.* Jöreskog and Sörbom 1979).

4. CONCLUDING REMARKS

The data sets generated from panel surveys are usually extremely rich in analytic potential. They contain repeated measures for some variables that are collected on several occasions, and also measures for other variables that are asked on a single wave. Repeated interviewing of the same sample provides the opportunity to collect data on new variables at each wave, thus yielding data on an extensive range of variables over a number of waves. A panel data set may be analyzed both longitudinally and cross-sectionally. Repeated measures may be used to examine individual response patterns over time, and they may also be related to other variables. Variables measured at a single wave may be analyzed both in relation to other variables measured at that wave and to variables measured at other waves.

The richness of panel data is of value only to the extent that the data set is analyzed, and analyzed in a timely manner. Running a panel survey is like being on a treadmill: the operations of questionnaire design, data collection, processing and analysis have to be undertaken repeatedly for each successive wave. There is a real danger that the survey team will become overwhelmed by this process, with the result that the data are not fully analyzed. To avoid this danger, adequate staffing is needed and a well-integrated organization needs to be established.

In addition it is advisable to keep the panel survey design simple. The survey design should be developed to meet clearly-specified objectives. Adding complexities to

the design to enhance the richness of the panel data set for other uses should be critically assessed. Although persuasive arguments can often be made for such additions, they should be rejected if they threaten the orderly conduct of any stage of the survey process.

As noted earlier, measurement errors have particularly harmful effects on the analysis of individual changes from panel survey data. The allocation of part of a panel survey's resources to measure the magnitude of such errors is therefore well warranted (Fuller 1989). Measurement errors may be investigated either by validity studies (comparing survey responses with "true" values from an external source) or by reliability studies (e.g., reinterview studies). The results of such studies may be then used in the survey estimation procedures to adjust for the effects of measurement errors.

REFERENCES

- ABOWD, H.M., and ZELLNER, A. (1985). Estimating gross flows. *Journal of Business and Economic Statistics*, 3, 254-283.
- BAILAR, B.A. (1975). The effects of rotation group bias on estimates from panel surveys. *Journal of the American Statistical Association*, 70, 23-30.
- BAILAR, B.A. (1989). Information needs, surveys, and measurement errors. *Panel Surveys*, (Eds. D. Kasprzyk, G. Duncan, G. Kalton and M.P. Singh). New York: John Wiley, 1-24.
- BINDER, D.A., and HIDIROGLOU, M.A. (1988). Sampling in time. *Handbook of Statistics*, (Vol. 6), (Eds. P.R. Krishnaiah and C.R. Rao). New York: North Holland, 187-211.
- BORUCH, R.F., and PEARSON, R.W. (1988). Assessing the quality of longitudinal surveys. *Evaluation Review*, 12, 3-58.
- BURGESS, R.D. (1989). Major issues and implications of tracing survey respondents. *Panel Surveys*, (Eds. D. Kasprzyk, G. Duncan, G. Kalton and M.P. Singh). New York: John Wiley, 52-74.
- CANTWELL, P.J., and ERNST, L.R. (1993). New developments in composite estimation for the Current Population Survey. *Proceedings: Symposium 92, Design and Analysis of Longitudinal Surveys, Statistics Canada*, 121-130.
- CHUA, T.C., and FULLER, W.A. (1987). A model for multinomial response error applied to labor flows. *Journal of the American Statistical Association*, 82, 46-51.
- CITRO, C.F., and KALTON, G. (1989). *Surveying the Nation's Scientists and Engineers*. Washington DC: National Academy Press.
- CITRO, C.F., and KALTON, G. (1993). *The Future of the Survey of Income and Program Participation*. Washington DC: National Academy Press.
- CLARRIDGE, B.R., SHEEHY, L.L., and HAUSER, T.S. (1978). Tracing members of a panel: a 17-year follow-up. *Sociological Methodology*, (Ed. K.F. Schuessler). San Francisco: Jossey-Bass, 389-437.
- CRIDER, D.M., WILLITS, F.K., and BEALER, R.C. (1971). Tracking respondents in longitudinal surveys. *Public Opinion Quarterly*, 35, 613-620.
- DOUGLAS, J.W.B. (1975). Early hospital admissions and later disturbances of behaviour and learning. *Developmental Medicine and Child Neurology*, 17, 456-480.
- DUNCAN, G.J., and KALTON, G. (1987). Issues of design and analysis of surveys across time. *International Statistical Review*, 55, 97-117.
- ECKLAND, B.K. (1968). Retrieving mobile cases in longitudinal surveys. *Public Opinion Quarterly*, 32, 51-64.
- EDWARDS, W.S., SPERRY, S., and EDWARDS, B. (1993). Using CAPI in a longitudinal survey: a report from the Medicare Current Beneficiary Survey. *Proceedings: Symposium 92, Design and Analysis of Longitudinal Surveys, Statistics Canada*, 21-30.
- ERNST, L.R. (1989). Weighting issues for longitudinal household and family estimates. In *Panel Surveys*, (Eds. D. Kasprzyk, G. Duncan, G. Kalton and M.P. Singh), New York: John Wiley, 139-159.
- FERBER, R. (1964). Does a panel operation increase the reliability of survey data: the case of consumer savings. *Proceedings of the Social Statistics Section, American Statistical Association*, 210-216.
- FULLER, W.A. (1989). Estimation of cross-sectional and change parameters: Discussion. *Panel Surveys*, (Eds. D. Kasprzyk, G. Duncan, G. Kalton and M.P. Singh). New York: John Wiley, 480-485.
- FULLER, W.A. (1990). Analysis of repeated surveys. *Survey Methodology*, 16, 167-180.
- FULLER, W.A., ADAM, A., and YANSANEH, I.S. (1993). Estimators for longitudinal surveys with application to the U.S. Current Population Survey. *Proceedings: Symposium 92, Design and Analysis of Longitudinal Surveys, Statistics Canada*, 309-324.
- GHANGURDE, P.D. (1982). Rotation group bias in the LFS estimates. *Survey Methodology*, 8, 86-101.
- GOLDSTEIN, H. (1979). *The Design and Analysis of Longitudinal Studies*. New York: Academic Press.
- HILL, D. (1987). Response errors around the seam: analysis of change in a panel with overlapping reference periods. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 210-215.
- HILL, M.S. (1992). *The Panel Study of Income Dynamics: A User's Guide*. Newbury Park, CA: Sage Publications.
- HINKINS, S., JONES, H., and SCHEUREN, F. (1988). Design modifications for the SOI corporate sample: Balancing multiple objectives. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 216-221.
- HSIAO, C. (1986). *Analysis of Panel Data*. New York: Cambridge University Press.
- HUANG, H. (1984). Obtaining cross-sectional estimates from a longitudinal survey: Experiences of the Income Survey Development Program. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 670-675.

- HUGGINS, V. (1987). Evaluation of missing wage data from the Survey of Income and Program Participation (SIPP). *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 205-209.
- JABINE, T.B., KING, K.E., and PETRONI, R.J. (1990). *Survey of Income and Program Participation Quality Profile*. Bureau of the Census, Washington DC: U.S. Department of Commerce.
- JÖRESKOG, K.G., and SÖRBOM, D. (1979). *Advances in Factor Analysis and Structural Equation Models*. Lanham MD: University Press of America.
- KALTON, G. (1986). Handling wave nonresponse in panel surveys. *Journal of Official Statistics*, 2, 303-314.
- KALTON, G., KASPRZYK, D., and McMILLEN, D.B. (1989). Nonsampling errors in panel surveys. *Panel Surveys*, (Eds. D. Kasprzyk, G. Duncan, G. Kalton and M.P. Singh). New York: John Wiley, 249-270.
- KALTON G., and LEPKOWSKI, J.M. (1985). Following rules in SIPP. *Journal of Economic and Social Measurement*, 13, 319-329.
- KALTON, G., and MILLER, M.E. (1991). The seam effect with Social Security income in the Survey of Income and Program Participation. *Journal of Official Statistics*, 7, 235-245.
- KASPRZYK, D. (1988). *The Survey of Income and Program Participation: An Overview and Discussion of Research Issues*. SIPP Working Paper No. 8830. Washington DC: U.S. Bureau of the Census.
- KASPRZYK, D., DUNCAN, G., KALTON, G., and SINGH, M.P. (Eds.) (1989). *Panel Surveys*. New York: John Wiley.
- KESSLER, R.C., and GREENBERG, D.F. (1981). *Linear Panel Analysis*. New York: Academic Press.
- KEYFITZ, N. (1951). Sampling with probabilities proportional to size: adjustment for changes in the probabilities. *Journal of the American Statistical Association*, 46, 183-201.
- KISH, L. (1965). *Survey Sampling*. New York: John Wiley.
- KISH, L., and SCOTT, A. (1971). Retaining units after changing strata and probabilities. *Journal of the American Statistical Association*, 66, 461-470.
- LAVALLÉE, P., and HUNTER, L. (1993). Weighting for the Survey of Labour and Income Dynamics. *Proceedings: Symposium 92, Design and Analysis of Longitudinal Surveys, Statistics Canada*, 65-75.
- LAZARSFELD, P.F. (1948). The use of panels in social research. *Proceedings of the American Philosophical Society*, 42, 405-410.
- LAZARSFELD, P.F., and FISKE, M. (1938). The panel as a new tool for measuring opinion. *Public Opinion Quarterly*, 2, 596-612.
- LEPKOWSKI, J.M. (1989). Treatment of wave nonresponse in panel surveys. *Panel Surveys*, (Eds. D. Kasprzyk, G. Duncan, G. Kalton and M.P. Singh). New York: John Wiley, 348-374.
- LEPKOWSKI, J.M., MILLER, D.P., KALTON, G., and SINGH, R. (1993). Imputation for wave nonresponse in the SIPP. *Proceedings: Symposium 92, Design and Analysis of Longitudinal Surveys, Statistics Canada*, 99-109.
- MAGNUSSON, D., and BERGMAN, L.R. (Eds.) (1990). *Data Quality in Longitudinal Research*. New York: Cambridge University Press.
- MARKUS, G.B. (1979). *Analyzing Panel Data*. Beverly Hills, CA: Sage Publications.
- MICHAUD, S., and HUNTER, L. (1993). Strategy for minimizing the impact of non-response for the Survey of Labour and Income Dynamics. *Proceedings: Symposium 92, Design and Analysis of Longitudinal Surveys, Statistics Canada*, 89-98.
- MOONEY, H.W. (1962). *Methodology in Two California Health Surveys*. Public Health Monograph No. 70, Washington DC: U.S. Department of Health, Education, and Welfare.
- MURRAY, T.S., MICHAUD, S., EGAN, M., and LEMAÎTRE, G. (1991). Invisible seams? The experience with the Canadian Labour Market Activity Survey. *Proceedings of the 1991 Annual Research Conference*. U.S. Bureau of the Census. Washington DC: U.S. Department of Commerce, 715-730.
- NELSON, D., McMILLEN, D., and KASPRZYK, D. (1985). *An Overview of the SIPP, Update I*. SIPP Working Paper No. 8401. Washington DC: U.S. Bureau of the Census.
- PENNELL, S.G., and LEPKOWSKI, J.M. (1992). Panel conditioning effects in the Survey of Income and Program Participation. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 566-571.
- RODGERS, W.L. (1989). Comparisons of alternative approaches to the estimation of simple causal models from panel data. *Panel Surveys*, (Eds. D. Kasprzyk, G. Duncan, G. Kalton and M.P. Singh). New York: John Wiley, 432-456.
- SINGH, R.P. (1993). Methodological experiments in the Survey of Income and Program Participation. *Proceedings: Symposium 92, Design and Analysis of Longitudinal Surveys, Statistics Canada*, 157-166.
- SINGH, R., HUGGINS, V., and KASPRZYK, D. (1990). *Handling Single Wave Nonresponse in Panel Surveys*. SIPP Working Paper No. 9009, Bureau of the Census, Washington DC: U.S. Department of Commerce.
- SKINNER, C.J. (1993). Logistic modelling of longitudinal survey data with measurement error. *Proceedings: Symposium 92, Design and Analysis of Longitudinal Surveys, Statistics Canada*, 269-276.
- SUBCOMMITTEE ON FEDERAL LONGITUDINAL SURVEYS (1986). *Federal Longitudinal Surveys*. Statistical Policy Working Paper 13. Washington DC: Office of Management and Budget.
- SUNTER, A.B. (1986). Implicit longitudinal files: A useful technique. *Journal of Official Statistics*, 2, 161-168.
- TRAUGOTT, M., and KATOSH, K. (1979). Response validity in surveys of voting behavior. *Public Opinion Quarterly*, 43, 359-377.
- U.S. BUREAU OF THE CENSUS (1978). *The Current Population Survey Design and Methodology*. Bureau of the Census Technical Paper No. 40, Washington DC: U.S. Government Printing Office.

VAN DE POL, F.J.R. (1989). *Issues of Design and Analysis of Panels*. Amsterdam: Sociometric Research Foundation.

VEROFF, J., HATCHETT, S., and DOUVAN, E. (1992). Consequences of participating in a longitudinal study of marriage. *Public Opinion Quarterly*, 56, 315-327.

WALL, W.D., and WILLIAMS, H.L. (1970). *Longitudinal Studies and the Social Sciences*. London: Heinemann.

WATERTON, J., and LIEVESLEY, D. (1989). Evidence of conditioning effects in the British Social Attitudes Panel. *Panel Surveys*, (Eds. D. Kasprzyk, G. Duncan, G. Kalton and M.P. Singh). New York: John Wiley, 319-339.

ACKNOWLEDGEMENTS

Survey Methodology wishes to thank the following persons who have served as referees, sometimes more than once, during 1993:

- | | |
|---|--|
| C.H. Alexander, <i>U.S. Bureau of the Census</i> | H. Lee, <i>Statistics Canada</i> |
| M.G. Arellano, <i>Advanced Linkage Technologies of America</i> | R.J.A. Little, <i>University of California – Los Angeles</i> |
| J. Armstrong, <i>Statistics Canada</i> | J. Liu, <i>Research Triangle Institute</i> |
| T.S. Arthanari, <i>Indian Statistical Institute</i> | D. Malec, <i>National Centers for Health Statistics</i> |
| M. Bankier, <i>Statistics Canada</i> | J.T. Massey, <i>U.S. Department of Health and Human Services</i> |
| Y. Beaucage, <i>Statistics Canada</i> | S.M. Miller, <i>U.S. Bureau of Labor Statistics</i> |
| T.R. Belin, <i>University of California – Los Angeles</i> | W.J. Mitofsky, <i>Voter Research and Surveys</i> |
| W. Bell, <i>U.S. Bureau of the Census</i> | G. Nathan, <i>Hebrew University</i> |
| D. Bellhouse, <i>University of Western Ontario</i> | H.B. Newcombe, <i>Consultant</i> |
| E. Berumen, <i>IASI</i> | C.A. Patrick, <i>Statistics Canada</i> |
| J. Bethel, <i>Westat</i> | D. Pfeffermann, <i>Hebrew University</i> |
| D.A. Binder, <i>Statistics Canada</i> | N.G.N. Prasad, <i>University of Alberta</i> |
| G.J. Brackstone, <i>Statistics Canada</i> | B. Quenneville, <i>Statistics Canada</i> |
| J.M. Brick, <i>Westat</i> | E. Rancourt, <i>Statistics Canada</i> |
| P.A. Chollette, <i>Statistics Canada</i> | J.N.K. Rao, <i>Carleton University</i> |
| G.H. Choudhry, <i>Statistics Canada</i> | L.-P. Rivest, <i>Université Laval</i> |
| A. Chu, <i>Westat</i> | L. Roy, <i>Statistics Canada</i> |
| M.J. Colledge, <i>Statistics Canada</i> | D. Royce, <i>Statistics Canada</i> |
| L. Cox, <i>U.S. Environmental Protection Agency</i> | D.B. Rubin, <i>Harvard University</i> |
| J.-C. Deville, <i>INSEE</i> | K. Rust, <i>Westat</i> |
| P. Dick, <i>Statistics Canada</i> | I. Sande, <i>Bell Communications Research</i> |
| J.D. Drew, <i>Statistics Canada</i> | C.-E. Särndal, <i>Université de Montréal</i> |
| R.E. Fay, <i>U.S. Bureau of the Census</i> | O. Sautory, <i>INSEE</i> |
| G. Forsman, <i>University of Linköping</i> | W.L. Schaible, <i>U.S. Bureau of Labor Statistics</i> |
| W.A. Fuller, <i>Iowa State University</i> | F.J. Scheuren, <i>U.S. Internal Revenue Service</i> |
| J.F. Gentleman, <i>Statistics Canada</i> | H.T. Schreuder, <i>U.S. Department of Agriculture</i> |
| M.E. Gonzalez, <i>U.S. Office of Management and Budget</i> | J. Sedransk, <i>State University of New York – Albany</i> |
| R.M. Groves, <i>U.S. Bureau of the Census</i> | C. Skinner, <i>University of Southampton</i> |
| M.A. Hidioglou, <i>Statistics Canada</i> | A. Singh, <i>Statistics Canada</i> |
| H. Hogan, <i>U.S. Bureau of the Census</i> | T.M.F. Smith, <i>University of Southampton</i> |
| G.J.C. Hole, <i>Statistics Canada</i> | N.L. Spruill, <i>U.S. Office of the Secretary of Defence</i> |
| D. Holt, <i>University of Southampton</i> | K.P. Srinath, <i>Statistics Canada</i> |
| D.G. Horvitz, <i>Retired</i> | C.M. Suchindran, <i>University of North Carolina – Chapel Hill</i> |
| D.A. Hubble, <i>U.S. Bureau of the Census</i> | S. Sudman, <i>University of Illinois – Urbana-Champaign</i> |
| A.Z. Israëls, <i>Netherlands Central Bureau of Statistics</i> | J.-L. Tambay, <i>Statistics Canada</i> |
| A.E. Johnson, <i>Agency for Health Care Policy and Research</i> | M.E. Thompson, <i>University of Waterloo</i> |
| G. Kalton, <i>Westat</i> | A.R. Tupek, <i>U.S. Bureau of Labor Statistics</i> |
| P.S. Kott, <i>U.S. Department of Agriculture</i> | J. Waksberg, <i>Westat</i> |
| P. Lahiri, <i>University of Nebraska – Lincoln</i> | W.E. Winkler, <i>U.S. Bureau of the Census</i> |
| D. Lalande, <i>Statistics Canada</i> | K.M. Wolter, <i>A.C. Nielsen</i> |
| P. Lavallée, <i>Statistics Canada</i> | J. Wretman, <i>Stockholm University</i> |
| L. Lebart, <i>CNRS – Télécom Paris</i> | A. Zaslavsky, <i>Harvard University</i> |

Acknowledgements are also due to those who assisted during the production of the 1993 issues: S. Beauchamp and S. Lineger (Photocomposition), and M. Haight (Translation Services). Finally we wish to acknowledge S. DiLoreto, M.M. Kent, C. Larabie and D. Lemire of Social Survey Methods Division, for their support with coordination, typing and copy editing.

Applied Statistics

JOURNAL OF THE ROYAL STATISTICAL SOCIETY (SERIES C)

CONTENTS

Volume 42, No. 3, 1993

	<i>Page</i>
Bayesian inference for generalized linear and proportional hazards models via Gibbs sampling <i>P. Dellaportas and A. F. M. Smith</i>	443
An investigation of changepoints in the annual number of cases of haemolytic uraemic syndrome <i>R. Henderson and J. N. S. Matthews</i>	461
Estimating annual total heron population counts <i>G. E. Thomas</i>	473
A bivariate ordered probit model with truncation: helmet use and motorcycle injuries <i>A. A. Weiss</i>	487
Modelling spatial patterns of trees attacked by bark-beetles <i>H. K. Preisler</i>	501
Calibration with many variables <i>M. C. Denham and P. J. Brown</i>	515
Attribute selection in correspondence analysis of incidence matrices <i>W. J. Krzanowski</i>	529
<i>General Interest Section</i>	
Analyses of public use decennial census data with multiply imputed industry and occupation codes <i>N. Schenker, D. J. Treiman and L. Weidman</i>	545
<i>Letter to the Editors</i>	557
<i>Statistical Software Reviews</i>	
MICLOG; PEST2.1	559
<i>Statistical Algorithms</i>	
AS 284 Null distribution of a statistic for testing sphericity and additivity: a Jacobi polynomial expansion <i>R. J. Boik</i>	567
AS 285 Multivariate normal probabilities of star-shaped regions <i>S. L. Lohr</i>	576
<i>Remark</i>	
AS R91 A remark on Algorithm AS 139: Maximum likelihood estimation in a linear model from confined and censored normal data <i>W. D. J. Ryder</i>	583

Printed in Great Britain at the Alden Press, Oxford

SPECIAL ISSUE

CONFIDENTIALITY AND DATA ACCESS

JOS 1993:2

Recent years have seen the unfortunate marriage of two issues of great concern to statistical agencies: confidentiality protection and steadily increasing nonresponse rates. To respond to these growing concerns, the Panel on Confidentiality and Data Access of the Committee on National Statistics and the Journal of Official Statistics have produced this special issue on Confidentiality and Data Access.

Contents

Preface

Report of the Panel on Confidentiality and Data Access

George T. Duncan, Virginia A. de Wolf, Thomas B. Jabine, and Miron L. Straf

Privacy and Advances in Social and Policy Sciences: Balancing Present Costs and Future Gains

Paul D. Reynolds

Measures of Disclosure Risk and Harm

Diane Lambert

Discussion: *Eleanor Singer, Thomas Plewis, L.W. Cook*

Informed Consent in U.S. Government Surveys

Robert H. Mugge

Informed Consent and Survey Response: A Summary of the Empirical Literature

Eleanor Singer

Discussion: *Tore Dalenius*

Masking Procedures for Microdata Disclosure Limitation

Wayne A. Fuller

Statistical Analysis of Masked Data

Roderick J.A. Little

Statistical Disclosure Limitation Practices of United States Statistical Agencies

Thomas B. Jabine

Discussion: *Brian V. Greenberg, Donald B. Rubin, Leon Willenborg*

Database Systems: Inferential Security

Sallie Keller-McNulty and Elizabeth A. Unger

Discussion: *Martin H. David, Gerald Gates, Teresa F. Lunt, Bo Sundgren*

Confidentiality Legislation and the United States Federal Statistical System

Joe S. Cecil

Procedures for Restricted Data Access

Thomas B. Jabine

Discussion: *Nancy J. Kirkendall, H.W. Watts, Photis Nanopoulos*

GUIDELINES FOR MANUSCRIPTS

Before having a manuscript typed for submission, please examine a recent issue (Vol. 10, No. 2 and onward) of Survey Methodology as a guide and note particularly the following points:

1. Layout

- 1.1 Manuscripts should be typed on white bond paper of standard size ($8\frac{1}{2} \times 11$ inch), one side only, entirely double spaced with margins of at least $1\frac{1}{2}$ inches on all sides.
- 1.2 The manuscripts should be divided into numbered sections with suitable verbal titles.
- 1.3 The name and address of each author should be given as a footnote on the first page of the manuscript.
- 1.4 Acknowledgements should appear at the end of the text.
- 1.5 Any appendix should be placed after the acknowledgements but before the list of references.

2. Abstract

The manuscript should begin with an abstract consisting of one paragraph followed by three to six key words. Avoid mathematical expressions in the abstract.

3. Style

- 3.1 Avoid footnotes, abbreviations, and acronyms.
- 3.2 Mathematical symbols will be italicized unless specified otherwise except for functional symbols such as “exp(·)” and “log(·)”, etc.
- 3.3 Short formulae should be left in the text but everything in the text should fit in single spacing. Long and important equations should be separated from the text and numbered consecutively with arabic numerals on the right if they are to be referred to later.
- 3.4 Write fractions in the text using a solidus.
- 3.5 Distinguish between ambiguous characters, (e.g., w, ω ; o, O, 0; l, 1).
- 3.6 Italics are used for emphasis. Indicate italics by underlining on the manuscript.

4. Figures and Tables

- 4.1 All figures and tables should be numbered consecutively with arabic numerals, with titles which are as nearly self explanatory as possible, at the bottom for figures and at the top for tables.
- 4.2 They should be put on separate pages with an indication of their appropriate placement in the text. (Normally they should appear near where they are first referred to).

5. References

- 5.1 References in the text should be cited with authors' names and the date of publication. If part of a reference is cited, indicate after the reference, e.g., Cochran (1977, p. 164).
- 5.2 The list of references at the end of the manuscript should be arranged alphabetically and for the same author chronologically. Distinguish publications of the same author in the same year by attaching a, b, c to the year of publication. Journal titles should not be abbreviated. Follow the same format used in recent issues.

Avant de dactylographier votre texte pour le soumettre, prière d'examiner un numéro récent de Techniques d'enquête (à partir du vol. 10, n° 2) et de noter les points suivants:

DIRECTIVES CONCERNANT LA PRÉSENTATION DES TEXTES

1. **Présentation**
 - 1.1 Les textes doivent être dactylographiés sur un papier blanc de format standard (8½ par 11 pouces), sur une face seulement, à double interligne partout et avec des marges d'au moins 1½ pouce tout autour.
 - 1.2 Les textes doivent être divisés en sections numérotées portant des titres appropriés.
 - 1.3 Le nom et l'adresse de chaque auteur doivent figurer dans une note au bas de la première page du texte.
 - 1.4 Les remerciements doivent paraître à la fin du texte.
 - 1.5 Toute annexe doit suivre les remerciements mais précéder la bibliographie.

2. **Résumé**

Le texte doit commencer par un résumé composé d'un paragraphe suivi de trois à six mots clés. Éviter les expressions mathématiques dans le résumé.

3. **Rédaction**
 - 3.1 Éviter les notes au bas des pages, les abréviations et les sigles.
 - 3.2 Les symboles mathématiques seront imprimés en italique à moins d'une indication contraire, sauf pour les symboles fonctionnels comme exp(·) et log(·) etc.
 - 3.3 Les formules courtes doivent figurer dans le texte principal, mais tous les caractères dans le texte doivent correspondre à un espace simple. Les équations longues et importantes doivent être séparées du texte principal et numérotées en ordre consécutif par un chiffre arabe à la droite si l'auteur y fait référence plus loin.
 - 3.4 Écrire les fractions dans le texte à l'aide d'une barre oblique.
 - 3.5 Distinguer clairement les caractères ambigus (comme w, ω; o, O; 0; 1, l).
 - 3.6 Les caractères italiques sont utilisés pour faire ressortir des mots. Indiquer ce qui doit être imprimé en italique en le soulignant dans le texte.

4. **Figures et tableaux**
 - 4.1 Les figures et les tableaux doivent tous être numérotés en ordre consécutif avec des chiffres arabes et porter un titre aussi explicatif que possible (au bas des figures et en haut des tableaux).
 - 4.2 Ils doivent paraître sur des pages séparées et porter une indication de l'endroit où ils doivent figurer dans le texte. (Normalement, ils doivent être insérés près du passage qui y fait référence pour la première fois.)

5. **Bibliographie**
 - 5.1 Les références à d'autres travaux faites dans le texte doivent préciser le nom des auteurs et la date de publication. Si une partie d'un document est citée, indiquer laquelle après la référence.
Exemple: Cochran (1977, p. 164).
 - 5.2 La bibliographie à la fin d'un texte doit être en ordre alphabétique et les titres d'un même auteur doivent être en ordre chronologique. Distinguer les publications d'un même auteur et d'une même année en ajoutant les lettres a, b, c, etc. à l'année de publication. Les titres de revues doivent être écrits au long. Suivre le modèle utilisé dans les numéros récents.

SPECIAL ISSUE

CONFIDENTIALITY AND DATA ACCESS

JOS 1993:2

Recent years have seen the unfortunate marriage of two issues of great concern to statistical agencies: confidentiality protection and steadily increasing nonresponse rates. To respond to these growing concerns, the Panel on Confidentiality and Data Access of the Committee on National Statistics and the Journal of Official Statistics have produced this special issue on Confidentiality and Data Access.

Contents

Preface	
Report of the Panel on Confidentiality and Data Access	George T. Duncan, Virginia A. de Wolf, Thomas B. Jabine, and Miron L. Sral
Privacy and Advances in Social and Policy Sciences: Balancing Present Costs and Future Gains	Paul D. Reynolds
Measures of Disclosure Risk and Harm	Diane Lambert
Discussion: Eleanor Singer, Thomas Plewis, L.W. Cook	
Informed Consent in U.S. Government Surveys	Robert H. Mudge
Informed Consent and Survey Response: A Summary of the Empirical Literature	Eleanor Singer
Discussion: Tore Dalenius	
Masking Procedures for Microdata Disclosure Limitation	Wayne A. Fuller
Statistical Analysis of Masked Data	Roderick J.A. Little
Statistical Disclosure Limitation Practices of United States Statistical Agencies	Thomas B. Jabine
Discussion: Brian V. Greenberg, Donald B. Rubin, Leon Willenborg	
Database Systems: Inferential Security	Sallie Keller-McNulty and Elizabeth A. Unger
Discussion: Martin H. David, Gerald Gates, Teresa F. Lunt, Bo Sundgren	
Confidentiality Legislation and the United States Federal Statistical System	Joe S. Cecil
Procedures for Restricted Data Access	Thomas B. Jabine
Discussion: Nancy J. Kirkendall, H.W. Watts, Phoeb Nanopoulos	

Applied Statistics

JOURNAL OF THE ROYAL STATISTICAL SOCIETY (SERIES C)

CONTENTS

Volume 42, No. 3, 1993

Page	
443	Bayesian inference for generalized linear and proportional hazards models via Gibbs sampling <i>P. Dellaportas and A. F. M. Smith</i>
461	An investigation of change-points in the annual number of cases of haemolytic uraemic syndrome <i>R. Henderson and J. N. S. Mathews</i>
473	Estimating annual total heron population counts <i>G. E. Thomas</i>
487	A bivariate ordered probit model with truncation: helmet use and motorcycle injuries <i>A. A. Weiss</i>
501	Modelling spatial patterns of trees attacked by bark-beetles <i>H. K. Preisler</i>
515	Calibration with many variables <i>M. C. Denham and P. J. Brown</i>
529	Attribute selection in correspondence analysis of incidence matrices <i>W. J. Krzanowski</i>
545	<i>General Interest Section</i> Analyses of public use decennial census data with multiply imputed industry and occupation codes <i>N. Schenker, D. J. Treiman and L. Weidman</i>
557	<i>Letter to the Editors</i>
559	<i>Statistical Software Reviews</i> MICLOG; PEST2.1
567	<i>Statistical Algorithms</i> AS 284 Null distribution of a statistic for testing sphericity and additivity: a Jacobi polynomial expansion <i>R. J. Boik</i>
576	AS 285 Multivariate normal probabilities of star-shaped regions <i>S. L. Lohr</i>
583	<i>Remark</i> AS R91 A remark on Algorithm AS 139: Maximum likelihood estimation in a linear model from confined and censored normal data <i>W. D. J. Ryder</i>

Printed in Great Britain at the Alden Press, Oxford

REMERCIEMENTS

Techniques d'enquête désire remercier les personnes suivantes, qui ont accepté de faire la critique d'un article, souvent plus d'une fois, durant l'année 1993:

C.H. Alexander, *U.S. Bureau of the Census*
M.G. Arellano, *Advanced Linkage Technologies of America*
J. Armstrong, *Statistique Canada*
T.S. Arthanari, *Indian Statistical Institute*
M. Bankier, *Statistique Canada*
Y. Beaucage, *Statistique Canada*
T.R. Belin, *University of California - Los Angeles*
W. Bell, *U.S. Bureau of the Census*
D. Bellhouse, *University of Western Ontario*
E. Berumen, *IASI*
J. Bethel, *Westat*
D.A. Binder, *Statistique Canada*
G.J. Brackstone, *Statistique Canada*
J.M. Brick, *Westat*
P.A. Cholleter, *Statistique Canada*
G.H. Choudhry, *Statistique Canada*
A. Chu, *Westat*
M.J. Colledge, *Statistique Canada*
L. Cox, *U.S. Environmental Protection Agency*
J.-C. Deville, *INSEE*
P. Dick, *Statistique Canada*
J.D. Drew, *Statistique Canada*
R.E. Fay, *U.S. Bureau of the Census*
G. Forsman, *University of Linköping*
W.A. Fuller, *Iowa State University*
J.F. Gentleman, *Statistique Canada*
M.E. Gonzalez, *U.S. Office of Management and Budget*
R.M. Groves, *U.S. Bureau of the Census*
M.A. Hidiroglou, *Statistique Canada*
H. Hogan, *U.S. Bureau of the Census*
G.J.C. Hole, *Statistique Canada*
D. Holt, *University of Southampton*
D.G. Horvitz, *Retraite*
D.A. Hubble, *U.S. Bureau of the Census*
A.Z. Israëls, *Netherlands Central Bureau of Statistics*
A.E. Johnson, *Agency for Health Care Policy and Research*
G. Kalton, *Westat*
P.S. Kott, *U.S. Department of Agriculture*
P. Lahiri, *University of Nebraska - Lincoln*
D. Lalonde, *Statistique Canada*
P. Lavallée, *Statistique Canada*
L. Lebart, *CNRS - Télécom Paris*

H. Lee, *Statistique Canada*
R.J.A. Little, *University of California - Los Angeles*
J. Liu, *Research Triangle Institute*
D. Malec, *National Centers for Health Statistics*
J.T. Massey, *U.S. Department of Health and Human Services*
S.M. Miller, *U.S. Bureau of Labor Statistics*
W.J. Mitofsky, *Voter Research and Surveys*
G. Nathan, *Hebrew University*
H.B. Newcombe, *Expert Conseil*
C.A. Patrick, *Statistique Canada*
D. Pfeffermann, *Hebrew University*
N.G.N. Prasad, *University of Alberta*
B. Quenneville, *Statistique Canada*
E. Rancourt, *Statistique Canada*
J.N.K. Rao, *Carleton University*
L.-P. Rivest, *Université Laval*
L. Roy, *Statistique Canada*
D. Royce, *Statistique Canada*
D.B. Rubin, *Harvard University*
K. Rust, *Westat*
I. Sande, *Bell Communications Research*
C.-E. Särndal, *Université de Montréal*
O. Sautory, *INSEE*
W.L. Schaible, *U.S. Bureau of Labor Statistics*
F.J. Scheuren, *U.S. Internal Revenue Service*
H.T. Schreuder, *U.S. Department of Agriculture*
J. Sedransk, *State University of New York - Albany*
C. Skinner, *University of Southampton*
A. Singh, *Statistique Canada*
T.M.F. Smith, *University of Southampton*
N.L. Spruill, *U.S. Office of the Secretary of Defence*
K.P. Srinath, *Statistique Canada*
C.M. Suchindran, *University of North Carolina - Chapel Hill*
S. Sudman, *University of Illinois - Urbana-Champaign*
J.-L. Tamabay, *Statistique Canada*
M.E. Thompson, *University of Waterloo*
A.R. Tupek, *U.S. Bureau of Labor Statistics*
J. Waksberg, *Westat*
W.E. Winkler, *U.S. Bureau of the Census*
K.M. Wolter, *A.C. Nielsen*
J. Wreiman, *Stockholm University*
A. Zaslavsky, *Harvard University*

On remercie également ceux qui ont contribué à la production des numéros de la revue pour 1991: S. Beauchamp et S. Linéger (Photocomposition), et M. Haigh (Services de traduction). Finalement on désire exprimer notre reconnaissance à S. DiLoreto, M.M. Kent, C. Larabie et D. Lémire de la Division des méthodes d'enquêtes sociales, pour leur apport à la coordination, la dactylographie et la rédaction.

- KALTON G., et LEPKOWSKI, J.M. (1985). Following rules in SIPP. *Journal of Economic and Social Measurement*, 13, 319-329.
- KALTON, G., et MILLER, M.E. (1991). The seam effect with Social Security income in the Survey of Income and Program Participation. *Journal of Official Statistics*, 7, 235-245.
- KASPRZYK, D. (1988). *The Survey of Income and Program Participation: An Overview and Discussion of Research Issues*. Document de travail du SIPP n° 8830. Washington DC: U.S. Bureau of the Census.
- KASPRZYK, D., DUNCAN, G., et SINGH, M.P. (Éds.) (1989). *Panel Surveys*. New York: John Wiley.
- KESSLER, R.C., et GREENBERG, D.F. (1981). *Linear Panel Analysis*. New York: Academic Press.
- KEYFITZ, N. (1951). Sampling with probabilities proportional to size: adjustment for changes in the probabilities. *Journal of the American Statistical Association*, 46, 183-201.
- KISH, L. (1965). *Survey Sampling*. New York: John Wiley.
- KISH, L., et SCOTT, A. (1971). Retaining units after changing strata and probabilities. *Journal of the American Statistical Association*, 66, 461-470.
- LAVALLÉE, P., et HUNTER, L. (1993). Méthodes de pondération pour l'enquête sur la dynamique du travail et du revenu. *Recueil: Symposium 92, Conception et analyse des enquêtes longitudinales, Statistique Canada*, 77-88.
- LAZARSFELD, P.F. (1948). The use of panels in social research. *Proceedings of the American Philosophical Society*, 42, 405-410.
- LAZARSFELD, P.F., et FISKE, M. (1938). The panel as a new tool for measuring opinion. *Public Opinion Quarterly*, 2, 596-612.
- LEPKOWSKI, J.M. (1989). Treatment of wave nonresponse in panel surveys. *Panel Surveys*, (Éds. D. Kasprzyk, G. Duncan, G. Kalton et M.P. Singh). New York: John Wiley, 348-374.
- LEPKOWSKI, J.M., MILLER, D.P., KALTON, G., et SINGH, R. (1993). Imputation pour la non-réponse de vague dans l'enquête Survey of Income and Program Participation (SIPP). *Recueil: Symposium 92, Conception et analyse des enquêtes longitudinales, Statistique Canada*, 115-126.
- MAGNUSSON, D., et BERGMAN, L.R. (Éds.) (1990). *Data Quality in Longitudinal Research*. New York: Cambridge University Press.
- MARKUS, G.B. (1979). *Analyzing Panel Data*. Beverly Hills, CA: Sage Publications.
- MICHAUD, S., et HUNTER, L. (1993). Stratégie pour minimiser l'impact de la non-réponse dans l'enquête sur la dynamique du travail et du revenu. *Recueil: Symposium 92, Conception et analyse des enquêtes longitudinales, Statistique Canada*, 103-114.
- MOONEY, H.W. (1962). *Methodology in Two California Health Surveys*. Public Health Monograph No. 70, Washington DC: U.S. Department of Health, Education, and Welfare.
- MURRAY, T.S., MICHAUD, S., EGAN, M., et LEMAITRE, G. (1991). Invisible seams? The experience with the Canadian Labour Market Activity Survey. *Proceedings of the 1991 Annual Research Conference*. U.S. Bureau of the Census, Washington DC: U.S. Department of Commerce, 715-730.
- NELSON, D., McMILLEN, D., et KASPRZYK, D. (1985). *An Overview of the SIPP, Update I*. SIPP Working Paper No. 8401. Washington DC: U.S. Bureau of the Census.
- PENNELL, S.G., et LEPKOWSKI, J.M. (1992). Panel conditioning effects in the Survey of Income and Program Participation. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 566-571.
- RODGERS, W.L. (1989). Comparisons of alternative approaches to the estimation of simple causal models from panel data. *Panel Surveys*, (Éds. D. Kasprzyk, G. Duncan, et M.P. Singh). New York: John Wiley, 432-456.
- SINGH, R.P. (1993). Expériences méthodologiques relatives à l'enquête Survey of Income and Program Participation. *Recueil: Symposium 92, Conception et analyse des enquêtes longitudinales, Statistique Canada*, 177-188.
- SINGH, R., HUGGINS, V., et KASPRZYK, D. (1990). *Handling Single Wave Nonresponse in Panel Surveys*. Document de travail du SIPP n° 9009, Bureau of the Census, Washington DC: U.S. Department of Commerce.
- SKINNER, C.J. (1993). Modélisation logistiqu de données d'enquête longitudinales pouvant comporter une erreur de mesure. *Recueil: Symposium 92, Conception et analyse des enquêtes longitudinales, Statistique Canada*, 301-309.
- SUBCOMMITTEE ON FEDERAL LONGITUDINAL SURVEYS (1986). *Federal Longitudinal Surveys*. Statistical Policy Working Paper 13. Washington DC: Office of Management and Budget.
- SUNTER, A.B. (1986). Implicit longitudinal files: A useful technique. *Journal of Official Statistics*, 2, 161-168.
- TRAUGOTT, M., et KATOSH, K. (1979). Response validity in surveys of voting behavior. *Public Opinion Quarterly*, 43, 359-377.
- U.S. BUREAU OF THE CENSUS (1978). *The Current Population Survey Design and Methodology*. Bureau of the Census Technical Paper No. 40, Washington DC: U.S. Government Printing Office.
- VAN DE POL, F.J.R. (1989). *Issues of Design and Analysis of Panels*. Amsterdam: Sociometric Research Foundation.
- VEROFF, J., HATCHETT, S., et DOUVAN, E. (1992). Consequences of participating in a longitudinal study of marriage. *Public Opinion Quarterly*, 56, 315-327.
- WALL, W.D., et WILLIAMS, H.L. (1970). *Longitudinal Studies and the Social Sciences*. London: Heinemann.
- WATERTON, J., et LIEVESLEY, D. (1989). Evidence of conditioning effects in the British Social Attitudes Panel. *Panel Surveys*, (Éds. D. Kasprzyk, G. Duncan, G. Kalton et M.P. Singh). New York: John Wiley, 319-339.

BIBLIOGRAPHIE

- ABOWD, H.M., et ZELTNER, A. (1985). Estimating gross flows. *Journal of Business and Economic Statistics*, 3, 254-283.
- BAILAR, B.A. (1975). The effects of rotation group bias on estimates from panel surveys. *Journal of the American Statistical Association*, 70, 23-30.
- BAILAR, B.A. (1989). Information needs, surveys, and measurement errors. *Panel Surveys*, (Eds. D. Kasprzyk, G. Duncan, et M.P. Singh). New York: John Wiley, 1-24.
- BINDER, D.A., et HIDIROGLOU, M.A. (1988). Sampling in time. *Handbook of Statistics*, (Vol. 6), (Eds. P.R. Krishnalah et C.R. Rao). New York: North Holland, 187-211.
- BORUCH, R.F., et PEARSON, R.W. (1988). Assessing the quality of longitudinal surveys. *Evaluation Review*, 12, 3-58.
- BURGESS, R.D. (1989). Major issues and implications of tracing survey respondents. *Panel Surveys*, (Eds. D. Kasprzyk, G. Duncan, et M.P. Singh). New York: John Wiley, 52-74.
- CANTWELL, P.J., et ERNST, L.R. (1993). Nouveaux développements dans l'estimation composite pour l'enquête Current Population Survey. *Recueil: Symposium 92, Conception et analyse des enquêtes longitudinales, Statistique Canada*, 139-149.
- CHUA, T.C., et FULLER, W.A. (1987). A model for multinomial response error applied to labor flows. *Journal of the American Statistical Association*, 82, 46-51.
- CITRO, C.F., et KALTON, G. (1989). *Surveying the Nation's Scientists and Engineers*. Washington DC: National Academy Press.
- CITRO, C.F., et KALTON, G. (1993). *The Future of the Survey of Income and Program Participation*. Washington DC: National Academy Press.
- CLARRIDGE, B.R., SHEEHY, L.L., et HAUSER, T.S. (1978). Tracing members of a panel: a 17-year follow-up. *Sociological Methodology*, (Ed. K.F. Schuessler). San Francisco: Jossey-Bass, 389-437.
- CRIDER, D.M., WILLITS, F.K., et BEALER, R.C. (1971). Tracking respondents in longitudinal surveys. *Public Opinion Quarterly*, 35, 613-620.
- DOUGLAS, J.W.B. (1975). Early hospital admissions and later disturbances of behaviour and learning. *Developmental Medicine and Child Neurology*, 17, 456-480.
- DUNCAN, G.J., et KALTON, G. (1987). Issues of design and analysis of surveys across time. *Revue Internationale de Statistique*, 55, 97-117.
- ECKLAND, B.K. (1968). Retrieving mobile cases in longitudinal surveys. *Public Opinion Quarterly*, 32, 51-64.
- EDWARDS, W.S., SPERRY, S., et EDWARDS, B. (1993). Utilisation de l'IPAO au cours d'une enquête longitudinale: un rapport sur l'enquête Medicare Current Beneficiary Survey. *Recueil: Symposium 92, Conception et analyse des enquêtes longitudinales, Statistique Canada*, 25-34.
- ERNST, L.R. (1989). Weighing issues for longitudinal household and family estimates. *Dans Panel Surveys*, (Eds. D. Kasprzyk, G. Duncan, et M.P. Singh). New York: John Wiley, 139-159.
- FERRBER, R. (1964). Does a panel operation increase the reliability of survey data: the case of consumer savings. *Proceedings of the Social Statistics Section, American Statistical Association*, 210-216.
- FULLER, W.A. (1989). Estimation of cross-sectional and change parameters: Discussion. *Panel Surveys*, (Eds. D. Kasprzyk, G. Duncan, et M.P. Singh). New York: John Wiley, 480-485.
- FULLER, W.A. (1990). Analyse d'enquêtes à passages répétés. *Techniques d'enquête*, 16, 177-190.
- FULLER, W.A., ADAM, A., et YANSANEH, I.S. (1993). Estimateurs pour des enquêtes longitudinales avec application à l'enquête Current Population Survey des E.-U. *Recueil: Symposium 92, Conception et analyse des enquêtes longitudinales, Statistique Canada*, 349-366.
- GHANGURDE, P.D. (1982). Le biais de renouvellement de l'échantillon dans les estimations de l'EPA. *Techniques d'enquête*, 8, 94-111.
- GOLDSTEIN, H. (1979). *The Design and Analysis of Longitudinal Studies*. New York: Academic Press.
- HILL, D. (1987). Response errors around the seam: analysis of change in a panel with overlapping reference periods. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 210-215.
- HILL, M.S. (1992). *The Panel Study of Income Dynamics: A User's Guide*. Newbury Park, CA: Sage Publications.
- HINKINS, S., JONES, H., et SCHEUREN, F. (1988). Design modifications for the SOI corporate sample: Balancing multiple objectives. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 216-221.
- HSIAO, C. (1986). *Analysis of Panel Data*. New York: Cambridge University Press.
- HUANG, H. (1984). Obtaining cross-sectional estimates from a longitudinal survey: Experiences of the Income Survey Development Program. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 670-675.
- HUGGINS, V. (1987). Evaluation of missing wage data from the Survey of Income and Program Participation (SIPP). *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 205-209.
- JABINE, T.B., KING, K.E., et PETRONI, R.J. (1990). *Survey of Income and Program Participation Quality Profile*. Bureau of the Census, Washington DC: U.S. Department of Commerce.
- JÖRESKOG, K.G., et SÖRBOM, D. (1979). *Advances in Factor Analysis and Structural Equation Models*. Lanham MD: University Press of America.
- KALTON, G. (1986). Handling wave nonresponse in panel surveys. *Journal of Official Statistics*, 2, 303-314.
- KALTON, G., KASPRZYK, D., et McMILLEN, D.B. (1989). Nonsampling errors in panel surveys. *Panel Surveys*, (Eds. D. Kasprzyk, G. Duncan, et M.P. Singh). New York: John Wiley, 249-270.

de mesure et, grâce à plusieurs mesures répétées, peut traiter des structures d'erreur corrélées (p. ex. Jöreskog et Sörbom 1979).

4. CONCLUSION

Les ensembles de données construits à partir d'enquêtes par panels offrent habituellement une grande richesse de contenu pour l'analyse. Ils contiennent des mesures répétées pour certaines variables observées plusieurs fois, et des mesures d'autres variables qui sont examinées à une seule vague. Les interviews répétées du même échantillon donnent l'occasion de recueillir des données sur de nouvelles variables à chaque vague, ce qui produit des données sur toute une gamme de variables observées sur plusieurs vagues. Les données d'un panel peuvent être soumises aussi bien à une analyse longitudinale qu'à une analyse transversale. Les mesures répétées peuvent permettre d'examiner les profils de réponse individuels dans le temps et peuvent aussi être reliées à d'autres variables. Les variables mesurées à une vague unique peuvent être analysées à la lumière d'autres variables observées à cette vague, ainsi que de variables mesurées à d'autres vagues.

La richesse des données d'un panel n'est mise à contribution que si les données sont effectivement analysées, et si elles le sont dans les délais les plus brefs possibles. Mettre en oeuvre une enquête par panel, c'est comme faire tourner une immense roue: les étapes de la conception du questionnaire, de la collecte des données, du traitement et de l'analyse doivent être reprises à chaque vague successive. Il y a un réel danger que l'équipe de l'enquête soit submergée par ce processus, et que les données ne soient pas pleinement analysées. Pour éviter un tel écueil, il faut mettre en place le personnel adéquat et construire une organisation bien intégrée.

Il est souhaitable, par ailleurs, de préserver la simplicité du plan de l'enquête par panel. Le plan devrait être élaboré à la lumière d'objectifs clairement définis. Un bon sens critique doit être exercé au moment d'ajouter au plan des éléments de complexité visant à accroître la richesse des données du panel pour en élargir les applications. Bien que des arguments persuasifs puissent souvent être invoqués en faveur de tels ajouts, ceux-ci doivent être rejetés s'ils menacent l'exécution ordonnée de n'importe quelle étape du processus d'enquête.

Comme il a été indiqué plus haut, les erreurs de mesure ont des effets particulièrement néfastes sur l'analyse des variations individuelles révélées par les données d'une enquête par panel. L'affectation d'une partie des ressources de l'enquête à la mesure de l'ampleur de ces erreurs est donc une décision judicieuse (Fuller 1989). Les erreurs de mesure peuvent être étudiées soit par des études de validité (c.-à-d. par une comparaison des réponses de l'enquête avec des valeurs "vraies" tirées d'une source externe), soit par des études de fiabilité (p. ex. études fondées sur des réinterviews). Les résultats de telles études peuvent ensuite être utilisés dans le processus d'estimation et permettre d'effectuer des rajustements tenant compte des erreurs de mesure.

à différents moments. Par exemple, à partir de données recueillies sur la cohorte britannique des naissances de 1946 dans le cadre de la National Survey of Health and Development, Douglas (1975) a constaté que les enfants qui étaient hospitalisés pendant plus d'une semaine ou qui subissaient des hospitalisations répétées entre les âges de 6 mois et de 3½ ans affichaient un comportement plus problématique à l'école et avaient des notes de lecture plus faibles à 15 ans. En principe, les enquêtes transversales peuvent utiliser des questions rétrospectives pour recueillir les données nécessaires à ce genre d'analyse. Cependant, les réponses à de telles questions sont souvent entachées par un niveau élevé d'erreur de mémoire et sont parfois empreintes de distorsions systématiques influant sur la relation étudiée.

• *Régression avec termes de différence.* Une régression dont les coefficients sont exprimés sous forme de différences peut permettre d'éviter un certain type d'erreurs dans la spécification d'un modèle. Supposons que le modèle de régression approprié pour la personne *i* au temps *t* soit

$$Y_{it} = \alpha + \beta x_{it} + \gamma z_{it} + \epsilon_{it},$$

où *x_{it}* est une variable explicative dont la valeur change avec le temps et *z_{it}* est une variable explicative qui est constante dans le temps (p. ex. sexe, race). Supposons en outre que *z_{it}* ne soit pas observée, par exemple qu'elle soit inconnue. Alors *β* peut encore être estimé, mais d'après une régression avec termes de différence:

$$Y_{it(t+1)} - Y_{it} = \beta (x_{it(t+1)} - x_{it}) + \epsilon_{it(t+1)} - \epsilon_{it},$$

(Roggers 1989; Duncan et Kalton 1987).

• *Estimation des durées des périodes.* Les données recueillies dans les enquêtes par panel peuvent servir à estimer la distribution des durées d'événements comme le fait de participer à un programme de protection sociale. Dans des enquêtes par panel comme l'enquête SIPP, certaines personnes se trouvent déjà dans une période de participation au début du panel (périodes tronquées au départ), certaines personnes amorcent une période de participation durant le panel, et certaines périodes se poursuivent après la fin du panel (périodes tronquées à droite). Par conséquent, les périodes ne sont pas toutes observées au complet. La distribution des durées des périodes peut être estimée en appliquant des méthodes d'analyse de survie, par exemple la méthode d'estimation de Kaplan-Meier fondée sur une limite de produit, à toutes les nouvelles périodes (y compris celles qui sont tronquées à droite) qui commencent pendant la durée du panel (p. ex. Ruggles et Williams 1989).

• *Modèles à équations structurelles avec erreurs de mesure.* La séquence de collecte des données dans une enquête par panel se traduit par une mise en ordre claire des variables de l'enquête, dont l'analyse se prête bien à l'utilisation de modèles à équations structurelles. Cette forme d'analyse peut prendre en considération les erreurs

effectuée auprès du panel de 1984 de la SIPP. À cette fin, une formule de vague manquante a servi à recueillir des réponses relatives à un sous-ensemble minimum de questions de base pour la vague manquante (Huggins 1987; Singh 1993). Les analyses ont montré que les données recueillies à l'aide de la formule de vague manquante affichaient un nombre passablement moindre de transitions que les données repères pour ce qui est de la réception de revenus, des biens et du recours à l'aide gouvernementale. Par conséquent, l'usage de la formule de vague manquante a été abandonné. Les dossiers administratifs peuvent parfois constituer une autre source de données de base pour les vagues manquantes.

3.2.2 Effet de conditionnement

Le biais dû au conditionnement, ou effet lié au temps passé dans l'échantillon, est l'effet qui s'exerce sur les réponses d'un membre du panel à une vague donnée de collecte des données en raison de sa participation aux vagues précédentes. L'effet peut refléter simplement un changement de comportement à l'égard de la déclaration. Par exemple, un répondant peut avoir remarqué dans une interview précédente que le fait de répondre "oui" à une question entraîne une série de sous-questions qui, en revanche, ne sont pas posées si la réponse est "non". Dans ce cas, le répondant peut répondre "non" pour éviter d'avoir à répondre à ces questions additionnelles. Du côté positif, un répondant peut avoir constaté à des interviews précédentes qu'il fallait fournir des renseignements détaillés sur le revenu et se préparer aux interviews suivantes en amassant au préalable l'information nécessaire. L'effet de conditionnement peut aussi se traduire par un changement du comportement réel. Par exemple, un répondant peut s'inscrire au programme des coupons alimentaires après avoir appris l'existence de ce programme grâce aux questions posées à ce sujet à des vagues précédentes de collecte des données.

Une récente étude expérimentale sur l'effet de conditionnement dans une enquête par panel de quatre ans sur les nouveaux mariés a permis de déceler un lien entre la participation à l'enquête et l'harmonie du mariage (Veroff et coll. 1992). Toutefois, cette étude s'est fondée sur des techniques d'interview en profondeur qui vont plus loin que celles utilisées dans la plupart des enquêtes. Un certain nombre d'études sur le conditionnement des panels ayant été effectuées dans des contextes d'enquête plus courants ont révélé que des effets de conditionnement se produisent parfois, mais qu'ils ne sont pas généralisés (Traugott et Katosh 1979; Ferber 1964; Mooney 1962; Waterton et Lienesky 1989). Les enquêtes par panel avec renouvellement et chevauchement ont comme avantage de permettre la comparaison entre des estimations relatives à la même période obtenues de panels différents. De telles comparaisons ont permis de détecter clairement la présence de ce qu'on appelle un "biais lié au groupe de renouvellement" dans les enquêtes sur la population active des États-Unis et du Canada (p. ex. Bailar 1975, 1989 et U.S. Bureau of the Census 1978 pour l'enquête "Current Population Survey" américaine;

Changurde 1982 pour l'enquête sur la population active canadienne). Le biais lié au groupe de renouvellement peut être le fruit d'un biais de non-réponse et d'effets de conditionnement. Dans des analyses comparant les panels chevauchants de 1985, 1986 et 1987 de l'enquête SIPP, Pennell et Lepkowski (1992) ont observé peu d'écart entre les résultats des différents panels.

3.2.3 Effet de lisière

Dans de nombreuses enquêtes par panel, des données sont recueillies relativement à des sous-périodes de la période de référence, à partir de la dernière vague de collecte de données. Dans l'enquête SIPP, par exemple, des données sont recueillies sur une base mensuelle à l'intérieur de la période de référence de quatre mois qui sépare les vagues. L'effet de lisière correspond à ce qu'on observe souvent avec cette forme de collecte de données, c'est-à-dire que les niveaux de changements déclarés entre des sous-périodes adjacentes (p. ex. adhérer à un programme social, puis cesser d'y participer d'un mois à l'autre) sont beaucoup plus élevés lorsque les données des deux sous-périodes proviennent de vagues différentes que lorsqu'elles viennent de la même vague. On a constaté dans l'enquête SIPP un très vaste effet de lisière, ayant trait aussi bien à l'état de prestataire qu'aux montants des prestations (voir, par exemple, Jabine et coll. 1990; Kalton et Miller 1991). L'effet de lisière s'est également manifesté dans l'enquête PSID (Hill 1987). Murray et coll. (1991) décrivent des méthodes permettant de réduire l'effet de lisière dans l'enquête sur la population active du Canada.

3.3 Analyse longitudinale

Il existe une documentation abondante, qui ne cesse de s'accroître, dans le domaine de l'analyse des données longitudinales, notamment plusieurs manuels qui en exposent la matière (p. ex. Goldstein 1979; Hsiao 1986; Kessler et Greenberg 1981; Markus 1979). La présente description ne peut être complète, et doit se limiter à quelques thèmes généraux.

- *Mesure de la variation brute.* Comme nous l'avons déjà signalé, un avantage essentiel de l'enquête par panel par rapport à l'enquête à passages répétés vient de la possibilité de mesurer la variation brute, c'est-à-dire la variation enregistrée au niveau individuel. La méthode de base pour mesurer la variation brute est de dresser un tableau comparatif des réponses fournies à une vague par rapport à celles fournies à la même question, lors d'une autre vague. Cette sorte d'analyse souffre toutefois d'une contrainte importante; en effet, les variations de l'erreur de mesure d'une vague à l'autre peuvent engendrer un biais prononcé dans l'estimation de la variation brute (pour un examen plus détaillé, voir Kalton et coll. 1989; Rodgers 1989; Abowd et Zellner 1985; Chua et Fuller 1987; Fuller 1990; et Skinner 1993).
- *Relation entre les variables dans le temps.* Les enquêtes par panel permettent de recueillir les données nécessaires à l'étude des liens qui existent entre des variables mesurées

ainsi que Lavallée et Hunter (1993), décrivent la méthode de pondération fondée sur des parts équitables qui peut être utilisée à cette fin.

● *Dépistage et suivi*. Une difficulté qui entrave la plupart

des enquêtes par panel vient du fait que certains membres du panel ont déménagé depuis la dernière vague et ne peuvent pas être retrouvés. Il y a deux façons de traiter ce problème. Premièrement, on peut tenter d'éviter qu'il ne survienne en mettant sur place un processus de suivi des membres du panel entre les vagues. Une méthode largement utilisée, lorsqu'il y a un long intervalle entre les vagues, consiste à faire des envois postaux aux répondants entre les vagues, par exemple à leur envoyer des cartes d'anniversaire ou des bulletins sur l'enquête, et à demander au service de la poste de fournir des avis de changement d'adresse, le cas échéant. Un autre moyen consiste à demander aux répondants d'indiquer les noms, les adresses et les numéros de téléphone de personnes qui leur sont proches (p. ex. des parents), qui sont peu susceptibles de déménager et qui pourront fournir leur adresse s'ils élisent domicile ailleurs.

La deuxième façon de traiter les cas de répondants introuvables est de mettre en oeuvre diverses méthodes de dépistage pour tenter de les retrouver. Avec des efforts et de l'ingéniosité, on peut atteindre des taux de réussite élevés. Certaines méthodes de dépistage peuvent être propres à la population à l'étude (p. ex. sociétés professionnelles pour les personnes faisant partie de groupes professionnels), tandis que d'autres peuvent être plus générales, par exemple: annuaire téléphonique, recherche informatisée de numéros de téléphones, annuaires téléphoniques inversés permettant d'obtenir les numéros de téléphone de voisins, suivi du courrier, registres des actes de mariage, immatriculation des véhicules, employeurs et services d'information financière. Il peut se révéler utile d'examiner les registres des décès, notamment dans le cas des enquêtes par panel de longue durée. Les membres du panel qui sont décédés peuvent ainsi être classés de façon appropriée, plutôt que d'être considérés comme des non-répondants. Les méthodes de dépistage sont examinées par Burgess (1989), Clarridge et coll. (1978), Cridder et coll. (1971) et Eckland (1968).

3.2 Problèmes des enquêtes par panel

Les enquêtes par panel sont confrontées, comme toutes les autres enquêtes, à une vaste gamme de sources d'erreur non due à l'échantillonnage. La présente section n'examine pas toutes ces sources, mais en présente trois qui sont propres aux enquêtes par panel: la non-réponse de vague, le biais de conditionnement et l'effet de lisière.

3.2.1 Non-réponse de vague

La non-réponse enregistrée à la première vague des enquêtes par panel correspond à celle qui s'observe dans les enquêtes transversales. Ce qui distingue les enquêtes par panel, c'est qu'elles font face à une non-réponse supplémentaire au moment de vagues subséquentes. Certains

membres du panel qui deviennent des non-répondants à une vague particulière ne répondent plus ensuite à aucune autre vague, tandis que d'autres répondent à nouveau à certaines ou à la totalité des vagues subséquentes. Les premiers sont souvent appelés des cas d'érosion, tandis que les seconds sont appelés des cas de non-érosion. Les taux globaux de non-réponse de vague des enquêtes par panel s'accroissent au fur et à mesure des vagues, mais dans le cas des enquêtes bien gérées, le taux d'augmentation s'atténue nettement avec le temps. Par exemple, dans le cas du panel de l'enquête SIPP de 1987, la perte était de 6,7% à la vague 1, de 12,6% à la vague 2, puis elle a augmenté lentement jusqu'à 19,0% à la vague 7 (Jabine et coll. 1990). La tendance de la non-réponse à s'atténuer aux vagues ultérieures est rassurante, mais le cumul de la non-réponse sur de nombreuses vagues produit néanmoins des taux de non-réponse élevés aux dernières vagues d'un panel de longue durée. Par exemple, en 1988, après 21 rondes annuelles de collecte des données, le taux de non-réponse de l'enquête PSID pour les personnes qui vivaient dans les ménages faisant partie de l'échantillon en 1968 avait atteint 43,9% (Hill 1992).

Le choix entre les deux méthodes d'usage général qui permettent de traiter les données d'enquête manquantes – pondération et imputation – n'est pas immédiat dans le cas de la non-réponse de vague touchant les enquêtes par panel. Pour l'analyse longitudinale, la méthode de pondération consiste à laisser de côté tous les enregistrés ayant une ou plusieurs vagues manquantes et à essayer de compenser pour leur absence par des rajustements de pondération appliqués aux enregistrés restants. Cette méthode peut entraîner une importante perte d'information lorsque le fichier de données porte sur plusieurs vagues. Dans la méthode d'imputation, en revanche, toutes les données déclarées sont conservées, mais on doit faire des imputations à vaste échelle à titre de compensation pour les vagues manquantes. Un compromis est aussi possible: on peut utiliser l'imputation pour certains profils de non-réponse de vague (p. ex. ceux qui ne comportent qu'une seule vague manquante, et pour lesquels on dispose des données des deux vagues adjacentes), et la pondération pour d'autres (voir, par exemple, Singh et coll. 1990). Pour l'analyse transversale, des fichiers de données distincts peuvent être créés à chaque vague. Ces fichiers peuvent comprendre tous les répondants à cette vague et utiliser soit des rajustements de pondération, soit des imputations, pour tenir compte des non-répondants à cette vague. Lepakowski (1986) et Lepakowski (1989) examinent les méthodes générales de traitement de la non-réponse de vague, Lepakowski et coll. (1993) traitent des imputations visant à compenser la non-réponse de vague dans la SIPP, et Michaud et Hunter (1993) décrivent les plans relatifs au traitement de la non-réponse de vague dans l'EDTR.

Dans le cas de la non-réponse de vague, il est possible de recueillir une partie ou la totalité des données de la vague manquante en effectuant une interview subséquente. Toutefois, la qualité des données rétrospectives recueillies de cette façon doit être évaluée avec soin. Une expérience visant à mesurer l'utilité de cette approche a été

déterminée en tenant compte du panel dans son ensemble et de toutes les vagues de collecte de données. En particulier, l'avantage d'une réduction des frais du travail sur place disparaît pour les vagues au cours desquelles l'interview téléphonique ou le questionnaire postal est utilisé. Par ailleurs, la migration de membres du panel vers des endroits se trouvant à l'extérieur des grappes originales réduit, aux vagues ultérieures, l'avantage que procurait la formation initiale des grappes en termes de réduction des coûts du travail sur place. (Toutefois, certains avantages des grappes initiales demeurent valables dans le cas de la proportion élevée des personnes mobiles qui déménagent à l'intérieur de leur propre quartier.) Le suréchantillonnage de certains sous-groupes de la population est largement utilisé dans les enquêtes transversales, afin de fournir un nombre suffisant de membres de ces sous-groupes pour permettre des analyses séparées. Voici des exemples de tels sous-groupes: personnes à faible revenu, minorités, groupe d'âge particulier, résidents d'un certain secteur géographique. Un tel suréchantillonnage peut aussi être utile aux enquêtes par panel, mais il faut se montrer prudent dans son application. Dans le cas des panels de longue durée, la prudence se justifie notamment par le fait que les objectifs de l'enquête peuvent changer avec le temps. Le suréchantillonnage visant à répondre à un objectif énoncé au début d'un panel peut se révéler nuisible aux objectifs qui s'imposeront plus tard. Un autre motif de prudence vient du fait que bon nombre des sous-groupes étudiés sont de nature transitoire (p. ex. personnes à faible revenu, personnes vivant dans un secteur géographique donné). Le suréchantillonnage de tels sous-groupes au début d'un panel pourrait avoir une valeur limitée lors des vagues ultérieures: certaines personnes incluses dans un tel suréchantillonnage sortiront du sous-groupe, tandis que d'autres exclues du suréchantillonnage y entreront. Troisièmement, la définition du sous-groupe doit être dans le contexte d'une analyse longitudinale doit être prise en considération. Par exemple, les données de l'enquête SIPP servent à estimer la durée des périodes de participation à divers programmes sociaux. Puisque ces estimations se fondent habituellement sur de nouvelles périodes qui commencent pendant la durée du panel, il pourrait être inutile de suréchantillonner les personnes déjà bénéficiaires de programmes sociaux. Voir Citro et Kaltou (1993) pour une analyse du suréchantillonnage dans le contexte de l'enquête SIPP.

Lorsque des données sur des personnes non membres de l'échantillon sont recueillies, il se peut que ces données servent simplement à décrire la situation des membres de l'échantillon, auquel cas les analyses se limitent aux membres de l'échantillon et les non-membres se voient attribuer un poids de zéro. Il est aussi possible d'inclure les personnes ne faisant pas partie de l'échantillon dans des analyses transversales. Dans ce cas, il faut élaborer une pondération appropriée des personnes appartenant et n'appartenant pas à l'échantillon, de façon à tenir compte des multiples façons dont les personnes peuvent figurer dans l'ensemble de données. Huang (1984), Ernst (1989),

Il importe également de prendre en considération les membres du panel qui sortent de la population à l'étude. Dans certains cas, le départ est irréversible (p. ex. décès), mais dans d'autres cas, il pourrait n'être que temporaire (p. ex. déménagement à l'étranger ou entrée en institution). Si l'on prend des mesures pour garder la trace des participants qui sortent du panel temporairement, ceux-ci peuvent être réadmis dès qu'ils réintègrent la population d'inférence.

Dans des enquêtes par panel comme l'enquête SIPP et la PSID, des données sont recueillies non seulement sur les membres des ménages de l'échantillon initial, mais également sur d'autres personnes – non membres de l'échantillon – avec lesquelles ils vivent lors de vagues ultérieures. L'objectif principal de la collecte de données d'enquête sur des personnes ne faisant pas partie de l'échantillon est de permettre de décrire la situation économique et sociale des membres de l'échantillon. Il convient, toutefois, de se demander si l'on devrait garder dans le panel une partie ou la totalité de ces personnes non membres de l'échantillon lorsqu'elles ne vivent plus avec les ménages du panel. Dans certains types d'analyse, il est utile de continuer de les suivre, mais cela oblige à leur consacrer une part importante des ressources de l'enquête.

Lorsque des données sur des personnes non membres de l'échantillon sont recueillies, il se peut que ces données servent simplement à décrire la situation des membres de l'échantillon, auquel cas les analyses se limitent aux membres de l'échantillon et les non-membres se voient attribuer un poids de zéro. Il est aussi possible d'inclure les personnes ne faisant pas partie de l'échantillon dans des analyses transversales. Dans ce cas, il faut élaborer une pondération appropriée des personnes appartenant et n'appartenant pas à l'échantillon, de façon à tenir compte des multiples façons dont les personnes peuvent figurer dans l'ensemble de données. Huang (1984), Ernst (1989),

Valeur du panel. Pour un niveau annuel donné de ressources, la taille d'échantillon de chaque panel est déterminée par les facteurs qui précèdent. On peut obtenir un panel de grande taille pour l'analyse longitudinale en allongeant la période de référence et en utilisant un plan sans chevauchement. Pour des estimations transversales, l'allongement de la période de référence peut permettre d'augmenter la taille de l'échantillon, mais le recours à un plan sans chevauchement n'a aucun effet semblable. La liste ci-dessus détermine les principaux paramètres d'un plan d'enquête par panel, mais il existe un certain nombre d'autres facteurs qui doivent aussi être examinés:

- **Mode de collecte des données.** Comme pour n'importe quelle enquête, il faut décider si les données seront recueillies par des interviews sur place, par téléphone ou par un questionnaire rempli par le répondant, et si l'on utilisera une méthode assistée par ordinateur (PAO) - interview sur place assistée par ordinateur ou ITAO - cas d'une enquête par panel, une telle décision doit être prise à l'égard de chaque vague de collecte des données, car on peut vouloir faire varier les modes de collecte (par exemple, interview sur place à la première vague pour établir un contact et créer un lien, et interviews par téléphone ou par questionnaire postal à certaines vagues subséquentes). Si le mode de collecte est susceptible de changer d'une vague à l'autre, il faut examiner la question de la comparabilité des données entre les vagues. Parfois, un changement de mode peut entraîner un changement d'interviewer, par exemple si l'on passe d'une interview sur place à une interview téléphonique assistée par ordinateur, effectuée à partir d'un endroit central. Dans un tel cas, l'effet du changement d'interviewer sur la volonté du répondant de continuer de participer au panel et sur la comparabilité des réponses entre les vagues doit aussi faire l'objet d'un examen attentif.
- **Interview avec rétro-information.** Dans les enquêtes par panel, il est possible de rappeler aux participants les réponses données à des vagues antérieures. Cette technique d'interview avec rétro-information permet

- *Plan d'échantillonnage.* La nature longitudinale d'une enquête par panel doit être prise en considération au moment de construire le plan d'échantillonnage visant la première vague. Un échantillonnage par grappes est souvent employé dans les enquêtes transversales avec interviews sur place, de façon à réduire les frais de déplacement et à limiter le travail de construction de la base de sondage à l'établissement de listes d'unités de logement pour certains segments seulement. Ces avantages sont obtenus au prix d'un d'accroissement de la variance des estimations de l'enquête, attribuable à l'utilisation de grappes. Le composition optimale des grappes dépend des divers coûts en cause et de l'homogénéité des variables de l'enquête à l'intérieur des grappes (voir, par exemple, Kish 1965). Dans le cas d'une enquête par panel, l'utilisation et l'ampleur des grappes devrait être

en raison aussi bien de l'érosion de l'échantillon que des difficultés de mise à jour de l'échantillon en fonction des nouveaux venus dans la population.

Il peut parfois être bénéfique de faire varier la durée du panel selon divers types de membres. Ainsi, lorsque les objectifs de l'analyse l'exigent, les membres du panel ayant certaines caractéristiques (p. ex. les membres d'une minorité) ou qui vivent certaines événements au cours de la durée du panel normal (p. ex. un divorce) peuvent être gardés dans le panel pour des périodes d'observation plus longues.

- *Durée de la période de référence.* La fréquence de collecte des données dépend de la capacité des répondants de se souvenir de l'information demandée. Ainsi, la PSID, qui comporte des vagues annuelles de collecte des données, exige des répondants qu'ils se souviennent des événements survenus au cours de l'année civile précédente, tandis que l'enquête SIPP, avec des vagues de collecte des données aux quatre mois, exige que l'on se souviennent de ce qui s'est passé au cours des quatre mois précédents. Plus la période de référence est longue, plus le risque d'erreur de mémoire est élevé.

- *Nombre de vagues.* Dans la plupart des cas, le nombre de vagues de collecte des données est déterminé par la durée du panel et la durée de la période de référence. Plus il y a de vagues, plus on risque de subir une érosion du panel et des effets de conditionnement, et plus lourd est le fardeau imposé aux répondants.

- *Panels chevauchants ou non chevauchants.* Dans le cas d'une enquête par panel à passages répétés de durée fixe, une décision doit être prise quant au chevauchement des panels. Prenons le cas, par exemple, d'une proposition d'un groupe d'étude du National Research Council selon lequel l'enquête SIPP devrait comporter un panel de quatre ans (Citro et Kalton 1993). Une possibilité est d'avoir un panel de quatre ans, et de commencer un nouveau panel dès que le précédent prend fin. On pourrait aussi établir des panels de quatre ans dont un nouveau serait amorcé tous les deux ans. Autre possibilité, on pourrait avoir des panels de quatre ans, mais en commencer un nouveau chaque année.
- Un plan comportant des panels non chevauchants a l'avantage d'être simple, car un seul panel à la fois est en cours de traitement. Il produit aussi un vaste échantillon pour l'analyse longitudinale; par exemple, les panels non chevauchants peuvent être approximativement deux fois plus gros que ceux d'un plan comportant en tout temps deux panels chevauchants. Toutefois, cet avantage d'une taille d'échantillon plus grande qu'offrent les panels non chevauchants ne vaut pas pour les estimations transversales, car s'il y a plusieurs panels simultanés, les données qui concernent un moment précis peuvent être combinées pour les besoins de l'estimation transversale. En outre, les estimations visant des données transversales observées vers la fin d'un panel, dans le cas d'un plan sans chevauchement, sont davantage susceptibles de souffrir d'un biais dû à l'érosion ou d'un biais

et Fiske 1938; Lazarsfeld 1948) et, depuis de nombreuses années, des enquêtes par panel sont réalisées dans de nombreux domaines. Les sujets suivants, par exemple, ont fait l'objet d'enquêtes par panel: croissance et développement humains, délinquance juvénile, consommation de drogues, victimisation, comportement des électeurs, études de marketing sur les dépenses de consommation, choix d'études et de carrières, retraite, santé, frais médicaux. (Voir Wall et Williams (1970) pour un examen d'études par panel réalisées il y a longtemps sur la croissance et le développement humains, Boruch et Pearson (1988) pour des descriptions de certaines enquêtes par panel menées aux États-Unis, et le Subcommittee on Federal Longitudinal Surveys (1986) pour des descriptions d'enquêtes par panel fédérales américaines.) On a pu constater ces dernières années un important gain d'intérêt pour les enquêtes par panel dans de nombreux secteurs, notamment dans le domaine de l'information économique sur les ménages. L'enquête permanente américaine "Panel Study of Income Dynamics" a commencé en 1968 (voir Hill 1992, pour une description de la PSID) et des enquêtes par panel de longue durée du même genre ont été amorcées dans de nombreux pays européens au cours de la dernière décennie. Le U.S. Bureau of the Census a lancé l'enquête "Survey of Income and Program Participation" (SIPP) en 1983 (Nelson et coll. 1985; Kasprzyk 1988; Jabin et coll. 1990), et Statistique Canada a introduit l'Enquête sur la dynamique du travail et du revenu (EDTR) en 1993. L'intérêt accru manifesté à l'égard des enquêtes par panel s'est aussi accompagné de la parution d'un nombre croissant d'ouvrages sur la méthodologie de telles enquêtes; citons, à titre d'exemples récents, Kasprzyk et coll. (1989), Magnusson et Bergman (1990) et Van de Pol (1989).

La présente section décrit les principaux aspects qui interviennent dans la conception et l'analyse des enquêtes par panel. Nous nous intéressons principalement aux enquêtes par panel à passages répétés de durée fixe, comme l'enquête SIPP et l'enquête EDRT, mais l'essentiel de notre examen s'applique, de façon générale, à toutes les formes d'enquêtes par panel.

3.1 Plan d'une enquête par panel: les choix à faire

La dimension "temps" est une dimension de complexité additionnelle que possède une enquête par panel par rapport à une enquête transversale. En sus de toutes les décisions qui doivent être prises dans le cadre de la planification d'une enquête transversale, une vaste gamme de choix additionnels doivent être faits pour une enquête par panel. Voici les principales décisions à prendre:

- *Durée du panel.* Plus le panel est de longue durée, plus grande est la richesse des données pour les fins de l'analyse longitudinale. Par exemple, plus le panel durera longtemps, plus il y aura de périodes de chômage commençant pendant la durée du panel qui se termineront avant la fin du panel, et donc plus l'estimation de la fonction de survie pour de telles périodes sera précise. En revanche, plus le panel est de longue durée, plus il est difficile de garder un échantillon transversal représentatif aux vagues finales du panel,

Les enquêtes à passages répétés permettent de recueillir des données sur des événements qui surviennent au cours d'une période donnée, ou encore sur la durée d'événements (p. ex. des périodes de maladie) au moyen de questions rétrospectives. Toutefois, les questions portant sur le passé, parce que les répondants éprouvent de la difficulté à se souvenir des dates, engendrent souvent un important problème d'erreur de réponse et font naître le risque d'un biais de télescopage. Une enquête par panel utilisant, pour les événements à étudier, une période de référence correspondant à l'intervalle entre les vagues de collecte des données peut éliminer le problème du télescopage, en utilisant l'interview précédente pour délimiter les événements (ainsi, on ne tiendra pas compte d'une maladie déclarée au cours de la présente interview si la même maladie avait été déclarée à l'interview précédente). De même, une enquête par panel peut permettre de déterminer la durée d'un événement d'après les vagues successives de collecte des données, en limitant la période de référence à l'intervalle entre les vagues.

Des collectes de données répétées au fil du temps peuvent être l'occasion de constituer progressivement un échantillon de membres d'une population rare, par exemple des personnes ayant une maladie chronique rare ou des personnes ayant récemment vécu un deuil. Les enquêtes à passages répétés peuvent être utilisées à cette fin et permettre de créer des échantillons de toutes sortes de populations rares. Les enquêtes par panel, toutefois, ne peuvent permettre d'accumuler que des événements rares nouveaux (p. ex. des deuils) et non des caractéristiques rares stables (p. ex. des personnes ayant une maladie chronique). Si un échantillon de membres ayant une caractéristique rare stable (p. ex. personnes ayant un doctorat) a déjà été établi, une enquête par panel peut être utile à la tenue à jour de cet échantillon, grâce à un apport approprié de nouveaux membres au moment de vagues ultérieures (voir, par exemple, Citro et Kalton 1989).

Les enquêtes par panel avec renouvellement visent principalement l'estimation des niveaux courants et de la variation nette (objectifs (a) et (c)). Dans de telles enquêtes, les éléments font généralement partie du panel pendant une courte période seulement. Par exemple, les membres de l'échantillon de l'enquête mensuelle sur la population active du Canada demeurent dans l'échantillon pendant une période de six mois seulement. Par conséquent, la mesure dans laquelle on peut évaluer les changements individuels et obtenir une agrégation des données dans le temps est limitée par la courte durée des panels. Une caractéristique spéciale des enquêtes par panel avec renouvellement réside dans la possibilité de recourir à l'estimation composée pour améliorer la précision tant de la variation nette (voir Binder et Hidiroglou 1988; Cantwell et Ernst 1993). Voir aussi Fuller et coll. (1993) pour une autre méthode d'utilisation de données antérieures dans le but de tirer des estimations d'un plan d'enquête par panel avec renouvellement. Comme les enquêtes par panel avec renouvellement, les enquêtes avec chevauchement visent principalement l'estimation des niveaux courants et de la variation nette. Elles

Lorsqu'on compare différents plans d'enquêtes longitudinales, les coûts doivent être pris en considération. Par exemple, les enquêtes par panel permettent d'éviter les coûts de la sélection répétée d'échantillons qu'exigent les enquêtes à passages répétés, mais comportent des coûts pour le dépistage et le suivi des déplacements des membres de l'échantillon, et parfois aussi pour l'offre d'incitatifs destinés à s'assurer la fidélité des membres du panel (voir la section 3). Si deux plans peuvent l'un et l'autre répondre aux objectifs de l'enquête, les coûts relatifs pour des niveaux de précision donnés des estimations de l'enquête doivent être examinés.

3. ENQUÊTES PAR PANEL

Les mesures répétées portant sur le même échantillon que permettent les enquêtes par panel confèrent à ces enquêtes un avantage clé, sur le plan de l'analyse, par rapport aux enquêtes à passages répétés. Les mesures de la variation brute et d'autres composantes de la variation individuelle qu'on peut extraire des données d'une enquête par panel sont la source d'une bien meilleure compréhension des processus sociaux que celle que peut procurer une série de clichés transversaux indépendants. Le potentiel des données longitudinales tirées des enquêtes par panel est depuis longtemps reconnu (voir, par exemple, Lazarsfeld

Un certain nombre de plans d'enquête ont été mis au point pour la collecte des données nécessaires à l'atteinte de ces objectifs. En voici une liste:

- *Enquêtes à passages répétées.* Une enquête à passages répétées est une série d'enquêtes transversales distinctes réalisées à différents moments. On ne cherche aucune-ment à s'assurer que les éléments de l'échantillon soient les mêmes d'un passage à l'autre. Les éléments de l'échantillon sont tirés d'une population définie de la même manière à chaque enquête (p. ex. mêmes frontières géographiques et limites d'âge) et les questions, pour une bonne part, sont identiques d'un passage à l'autre.
- *Enquêtes par panel.* Une enquête par panel consiste à recueillir des données, à différents moments, auprès du même échantillon de répondants.

• *Enquête par panel à passages répétées.* Une enquête par panel à passages répétées est formée d'une série d'enquêtes par panel ayant chacune une durée fixe. Il peut n'y avoir aucun chevauchement des périodes visées par les différents panels, par exemple si un panel ne commence qu'au moment où le précédent se termine (ou plus tard), ou encore il peut y avoir un chevauchement, si deux panels ou plus couvrent partiellement la même période.

• *Enquête par panel avec renouvellement.* À strictement parler, une enquête par panel avec renouvellement est équivalente à une enquête par panel à passages répétées dans laquelle il y a un chevauchement. Dans les deux cas, la durée d'un panel est limitée, et deux panels ou plus sont en cours d'examen au même moment. Toutefois, il est utile de faire une distinction entre ces deux plans, car ils ne visent pas les mêmes objectifs. Les enquêtes par panel avec renouvellement sont largement utilisées pour produire une série d'estimations transversales et d'estimations de variations nettes (p. ex. taux de chômage et évolution de ces taux), tandis que les enquêtes par panel à passages répétées avec chevauchement accordent en outre beaucoup d'importance à des mesures longitudinales (p. ex. durée des périodes de chômage). Par conséquent les enquêtes par panel à passages répétées ont tendance à être de plus longue durée et à comporter moins de panels en cours de traitement à un moment quelconque que les enquêtes par panel avec renouvellement.

• *Enquête avec chevauchement.* Comme une enquête à passages répétées, une enquête avec chevauchement est formée d'une série d'enquêtes transversales menées à différents moments. Toutefois, tandis que l'enquête à passages répétées ne cherche pas à obtenir un chevauchement des échantillons de l'enquête entre deux moments successifs, une enquête avec chevauchement vise précisément à en obtenir un. Le but pourrait être de maximiser le degré de chevauchement des échantillons tout en tenant compte à la fois des variations désirées des probabilités de sélection des éléments de l'échantillon qui demeurent dans la population de l'enquête et des variations de la composition de la population au fil du temps.

- *Enquête à panel fractionné.* Une enquête à panel fractionné est la combinaison d'une enquête par panel et d'une enquête à passages répétées ou d'une enquête par panel avec renouvellement.

Le choix du plan dans une situation particulière dépend des objectifs que l'on veut atteindre. Certains plans sont supérieurs vis-à-vis de certains objectifs, mais moins performants à d'autres égards. Il y a des plans qui ne satisfont aucunement à certains objectifs. Pour un examen détaillé, voir Duncan et Kaltton (1987).

Le principal atout d'une enquête à passages répétées est qu'elle comporte le prélèvement d'un nouvel échantillon à chaque passage, de sorte que chaque enquête transversale se fonde sur un échantillon probabiliste de la population qui existe à ce moment précis. Une enquête par panel se fonde sur un échantillon tiré de la population qui existait au début du panel. Bien que, parfois, des efforts soient faits pour ajouter à un panel des échantillons de nouveaux membres à des stades ultérieurs, une telle mise à jour se révèle généralement difficile et est empreinte d'imperfections. De plus, les pertes dues à la non-réponse qui sont enregistrées à mesure qu'un panel vieillit accentuent le problème du biais de non-réponse dont sont entachées, à des points ultérieurs, les estimations de paramètres transversaux tirées du panel. Pour ces raisons, les enquêtes à passages répétées sont supérieures aux enquêtes par panel pour produire des estimations transversales et des moyennes de ces dernières (objectifs (a) et (b)). Dans le cas des moyennes d'estimations transversales, un autre facteur à prendre en considération est la corrélation qui existe entre les valeurs des variables de l'enquête pour la même personne à différents moments. Si cette corrélation est positive, comme c'est le cas généralement, il en résulte un accroissement des erreurs-types des moyennes des estimations transversales provenant d'une enquête par panel. Ce facteur joue donc aussi en faveur des enquêtes à passages répétées par rapport aux enquêtes par panel, lorsqu'on s'intéresse à des moyennes d'estimations transversales.

La représentation plus fidèle des échantillons d'une enquête à passages répétées à des stades ultérieurs semblerait aussi favoriser ce type d'enquêtes par rapport aux enquêtes par panel pour l'estimation de la variation nette (en supposant qu'on veut mesurer cette dernière en tenant compte des changements touchant aussi bien la composition que les caractéristiques de la population). Toutefois, dans ce cas, les corrélations positives des valeurs des variables de l'enquête pour la même personne au fil du temps ont pour effet de diminuer les erreurs-types des estimations de la variation nette établie d'après une enquête par panel. Par conséquent, la présence de ces corrélations joue en faveur de l'enquête par panel pour la mesure de la variation nette.

Les avantages principaux des enquêtes par panel résident dans leur capacité de mesurer la variation brute, ainsi que de permettre l'agrégation, au fil du temps, de données sur les membres de l'échantillon (objectifs (d) et (e)). Les enquêtes à passages répétées ne peuvent répondre à ces objectifs. L'énorme potentiel d'analyse offert par la mesure de variations touchant les mêmes personnes est la principale raison incitant à recourir à une enquête par panel.

Enquêtes par panel: ajout d'une quatrième dimension

GRAHAM KALTON et CONSTANCE F. CITRO¹

RÉSUMÉ

Les enquêtes qui consistent à recueillir des données dans le temps peuvent viser de nombreux objectifs. Dans la première moitié du présent article, nous examinons diverses options de plans d'enquête – enquêtes à passages répétés, enquêtes par panel, enquêtes par panel avec renouvellement et enquêtes à panel fractionné – pouvant permettre d'atteindre ces objectifs. La deuxième moitié est axée sur les enquêtes par panel. Nous y traitons des décisions qui doivent être prises au moment de la conception d'une enquête par panel, des problèmes posés par la non-réponse aux différentes vagues, du biais de conditionnement et de l'effet de lisière, ainsi que de certaines méthodes permettant l'analyse longitudinale des données d'enquête par panel.

MOTS CLÉS: Enquêtes par panel; enquêtes par panel avec renouvellement; enquêtes à passages répétés; érosion du panel; biais de conditionnement; effet de lisière; analyse longitudinale.

1. INTRODUCTION

Les populations visées par les enquêtes changent constamment dans le temps, aussi bien parce que leur composition évolue qu'en raison de modifications des caractéristiques de leurs membres. Des changements de composition surviennent quand des membres entrent dans la population au moment de la naissance (ou de l'atteinte de l'âge adulte), ou encore lorsqu'ils immigreront ou qu'ils sortent d'une institution (si la population étudiée ne comprend pas les personnes en institution); de même, les décès, l'émigration et l'entrée en institution sont des causes de changement. Les caractéristiques changent, par exemple, lorsque des personnes mariées divorcent, ou lorsque le revenu mensuel d'une personne passe de \$2 000 à \$2 500. Ces changements touchant la population sont à l'origine de toute une série d'objectifs des analyses de l'évolution des données d'enquête en fonction du temps. La présente communication passe en revue les plans d'enquête qui produisent les données nécessaires à l'atteinte de ces divers objectifs.

Cette communication se divise en deux parties. La première présente les aspects généraux de la réalisation d'enquêtes longitudinales, notamment les objectifs de ces enquêtes et les types de plans d'enquête possibles. Cet examen fait l'objet de la section 2. La deuxième partie, qui est le corps principal de la présente communication, s'intéresse au cas particulier des enquêtes par panel, qui consistent à suivre le même échantillon d'unités au fil du temps. Les différents aspects de la conception, de l'exécution et de l'analyse d'une enquête par panel sont exposés à la section 3. La section 4 présente quelques remarques en guise de conclusion.

2. ENQUÊTES LONGITUDINALES

La présente section décrit de façon générale les objectifs analytiques propres aux données qui évoluent dans le

temps, les plans d'enquête qui conviennent à ces données et la mesure dans laquelle les différents plans peuvent satisfaire aux divers objectifs. L'exposé s'inspire fortement de celui de Duncan et Kalton (1987), qui contiennent un traitement plus détaillé de ces questions.

L'évolution des caractéristiques et de la composition de la population dans le temps est à la source de toute une gamme d'objectifs ayant trait aux enquêtes longitudinales. Ces objectifs sont, notamment:

- (a) l'estimation des paramètres de la population (p. ex. la proportion de la population vivant sous le seuil de la pauvreté) à divers points dans le temps;
- (b) l'estimation des valeurs moyennes de paramètres de la population dans le temps (p. ex. moyenne annuelle de la consommation quotidienne de fer);
- (c) l'estimation de variations nettes, c'est-à-dire de variations à des niveaux d'aggrégation élevés (p. ex. variation de la proportion de chômeurs d'un mois à l'autre);
- (d) l'estimation de variations brutes et d'autres facettes de changements touchant les individus (p. ex. proportion de personnes qui étaient sous le seuil de la pauvreté une année et qui ne l'étaient plus l'année suivante);
- (e) l'aggrégation sur une certaine période de données relatives aux individus (p. ex. sommation des revenus de douze mois pour obtenir le revenu annuel);
- (f) l'obtention de données sur des événements qui se produisent au cours d'une période donnée (p. ex. tomber en chômage) et sur leurs caractéristiques (p. ex. durée des périodes de chômage);
- (g) l'accumulation progressive d'échantillons, notamment des échantillons de populations rares (p. ex. les femmes qui deviennent veuves);
- (h) la tenue à jour d'un échantillon de membres d'une population rare observée à un moment particulier (p. ex. scientifiques et ingénieurs relevés dans une enquête à grande échelle à un moment donné).

¹ Graham Kalton, Westat, 1650 Research Blvd., Rockville (Maryland), E.-U., 20850; Constance F. Citro, National Research Council, 2101 Constitution Ave., N.W., Washington, D.C., E.-U., 20418.

BIBLIOGRAPHIE

- COCHRAN, W.G. (1977). *Sampling Techniques*, (3^e Ed.). New York: John Wiley.
- CUNIA, T. (1965). Continuous forest inventory, partial replacement of samples and multiple regression. *Forest Science*, 11, 480-502.
- GROSENBAGH, L.R., et STOVER, W.S. (1957). Point-sampling compared with plot-sampling in southeast Texas. *Forest Science*, 3, 2-14.
- HUSCH, B. (1955). Results of an investigation of the variable plot method of cruising. *Journal of Forestry*, 53, 570-574.
- KISH, L. (1965). *Survey Sampling*. New York: John Wiley.
- ODERWALD, R.G. (1981). Point and plot sampling – the relationship. *Journal of Forestry*, 79, 377-378.
- PALLEY, M.N., et HORWITZ, L.G. (1961). Properties of some random and systematic point sampling estimators. *Forest Science*, 7, 52-65.
- ROESCH, F.A. Jr. (1993). Adaptive cluster sampling for forest inventories. *Forest Science*, 39, À paraître.
- ROESCH, F.A. Jr., GREEN, E.J., et SCOTT, C.T. (1989). New compatible estimators for survivor growth and ingrowth from remeasured horizontal point samples. *Forest Science*, 35, 281-293.
- ROESCH, F.A. Jr., GREEN, E.J., et SCOTT C.T. (1991). Compatible basal area and number of trees estimators from remeasured horizontal point samples. *Forest Science*, 37, 136-145.
- ROESCH, F.A. Jr., GREEN, E.J., et SCOTT, C.T. (1993). A test of alternative estimators for volume at time 1 from remeasured point samples. *Canadian Journal of Forest Research*, 23, 598-604.
- SCHREUDER, H.T. (1970). Point sampling theory in the framework of equal-probability cluster sampling. *Forest Science*, 16, 240-246.
- VAN DEUSEN, P.C., DELL, T.R., et THOMAS, C.E. (1986). Volume growth estimation from permanent horizontal points. *Forest Science*, 32, 415-422.

4. CONCLUSION

Nous avons présenté un concept d'échantillonnage forestier généralisé qui utilise un nombre fini de segments de terrain à titre d'échantillonnage d'une base formée d'une zone terrestre. Nous avons aussi présenté des estimateurs fondés sur ce concept. Le concept du casse-tête devrait aider à comprendre les similitudes et les différences entre différentes méthodes d'échantillonnage des forêts, en intégrant toutes les méthodes au même cadre de référence. S'il est vrai que nous n'utiliserions pas normalement les estimateurs dans la forme présentée pour réaliser un échantillonnage forestier réel, nous pouvons toujours trouver une forme calculable équivalente. L'avantage additionnel de cette approche différente pour les simulations d'échantillonnage n'est pas seulement de nature théorique, mais aussi de nature économique. Compte tenu du temps et de l'argent qu'il faut consacrer à la collecte des données dans les études de forestier, la capacité de tester facilement les propriétés de différentes méthodes d'échantillonnage avant de les mettre en oeuvre est d'une importance capitale. Nous ne voulons d'aucune façon sous-estimer la nécessité d'effectuer au départ un développement théorique rigoureux des plans d'échantillonnage forestier proposés, mais la simulation de ces plans avant leur mise en oeuvre peut aider à révéler des problèmes jusqu'à la non détection. Cette nouvelle approche facilitera de façon générale les comparaisons à l'intérieur de n'importe quel groupe de plans d'échantillonnage des forêts.

REMERCIEMENTS

Les auteurs tiennent à remercier le rédacteur associé, deux arbitres anonymes, ainsi que Hans Schreuder, pour leurs commentaires utiles.

mesures. Cette notion d'unité d'échantillonnage est utile pour comprendre les estimateurs des composantes du changement entre le temps 1 et le temps 2 donnés dans Van Deusen et coll. (1986) et Roesch et coll. (1989 et 1991).

3. ANALYSE

Compte tenu de la simplicité du concept du casse-tête, on peut se demander pourquoi cette méthode d'échantillonnage des forêts n'a pas été proposée avant. La raison la plus évidente est probablement que les estimateurs ci-dessus ne peuvent être calculés si les A_j sont inconnues. Puisque la zone de sélection d'un arbre particulier peut être répartie entre de nombreuses pièces du casse-tête et que la taille d'une pièce particulière du casse-tête peut être limitée par des arbres non sélectionnés à partir de cette pièce, les zones de sélection aussi bien des arbres de l'échantillon que des autres arbres doivent être connues pour qu'on puisse calculer les A_j des segments sélectionnés. Par exemple, à la figure 1, si notre point tombait dans la section c, nous sélectionnerions les arbres 1 et 2, de sorte que la superficie c + d pourrait être directement calculée. Toutefois, pour calculer X et $v(X)$, nous devons connaître la superficie du segment c seul, et l'échantillonnage ne nous donne pas suffisamment d'information à cette fin. Nous allons voir que cette lacune apparente n'est pas importante, en montrant que X peut être reformulé en des termes qui sont calculables. Ce sera, en fait, toujours le cas, peu importe à quelle méthode d'échantillonnage le concept du casse-tête sera appliqué.

Le concept du casse-tête, dans le cas de l'échantillonnage par points, revient à appliquer la population des arbres sur la population connexe des segments de terrain. Nous pouvons reformuler X afin de montrer que ce dernier équivaut à l'estimateur d'échantillonnage par points habituel fondé sur la population des arbres. En développant l'équation (2) de manière à inclure la définition de y_j , et après quelques réagencements, on obtient:

$$Y = \overline{A_T} \sum_{j=1}^m \frac{A_j}{y_j} w_j$$

$$\overline{A_T} = \frac{\sum_{j=1}^m \frac{A_j}{p_{ij} y_i}}{w_j}$$

$$\overline{A_T} = \sum_M^m \sum_N^j \frac{A_j}{y_i z_{ij} w_j}$$

$$\overline{A_T} = \sum_N^m \sum_M^j \frac{A_j}{y_i} z_{ij} w_j$$

(10)

$$\overline{A_T} = \sum_N^m \frac{A_j}{y_i} w_i,$$

où w_i est égal au nombre de fois que l'arbre i est prélevé dans l'échantillon. L'expression finale dans (10) est l'estimateur de l'échantillonnage par points habituel.

La présente communication n'a pas pour but, par conséquent, de présenter un nouvel ensemble d'estimateurs pour des systèmes d'échantillonnage qui comptent déjà des estimateurs raisonnablement bons, mais plutôt de montrer comment des plans d'échantillonnage qui font l'objet de justifications très disparates dans la littérature se rejoignent à un niveau général. Cette perception différente peut être utile de multiples façons. Nous croyons, tout d'abord, que certains systèmes abstraits d'échantillonnage des forêts pourraient être plus faciles à comprendre s'ils étaient adaptés au cadre décrit ci-dessus. Nous avons constaté, par exemple, que les étudiants saisissent facilement la notion d'échantillonnage par points quand celui-ci est décrit comme une simple méthode de découpage de la forêt en pièces de casse-tête sans chevauchement, qui sont ensuite soumises à un échantillonnage avec probabilité proportionnelle à la taille. Les chercheurs intéressés à élaborer de nouveaux plans d'échantillonnage pour les forêts, ou de nouveaux estimateurs pour les plans existants, peuvent tirer parti de ce concept, car celui-ci leur offre une autre façon de percevoir leurs nouveaux plans et de programmer les simulations d'échantillonnage forestier qui serviront à tester les nouvelles méthodes. La simulation analysée dans Roesch (1993), par exemple, a été simplifiée par l'utilisation du concept du casse-tête plutôt que les autres concepts de base d'échantillonnage forestier qui avaient été proposés jusque-là. La simplification est venue du fait que l'essentiel de la simulation pouvait servir à de nombreux plans d'échantillonnage différents, seules quelques modifications minimales étant apportées au sous-programme de découpage du casse-tête.

Puisque les simulations d'échantillonnage forestier se fondent souvent sur une forêt cartographiée, les A_j peuvent être obtenues directement. Une fois le casse-tête découpé, y_j peut être calculé pour chaque pièce. Puis le programme de simulation choisit simplement les pièces à partir d'une liste, avec probabilité proportionnelle à la taille. Par conséquent, voyons ce qu'il en est d'une simulation fondée sur l'utilisation du point comme unité d'échantillonnage. Dans ce cas, un point est sélectionné au hasard, et le programme parcourt la liste des arbres pour trouver tous ceux qui sont suffisamment près de ce point pour être inclus dans l'échantillon. Les attributs d'intérêt sont ensuite calculés. Puisque la probabilité de choisir le même point deux fois pour une population infinie est nulle, cette recherche dans la liste et ce calcul devraient être repris pour chaque point choisi au hasard, et entraîneraient peut-être un calcul répété des attributs pour la même grappe d'arbres. Pour les fins des simulations, la méthode optimale de programmation dépendra de la longueur de la liste d'arbres à parcourir, de la répartition en grappes dans la population des arbres et du nombre de points choisis au hasard.

Nous pouvons maintenant montrer que Y est non biaisé pour Y en montrant que $Y^* = Y$. En substituant le côté droit de l'équation (1) à y_j dans la définition de Y^* , nous obtenons:

$$Y^* = \sum_{N=1}^M \sum_{j=1}^I p_{ij} y_i. \quad (5)$$

En insérant la définition de p_{ij} et en réorganisant l'ordre de sommation, nous obtenons:

$$Y^* = \sum_{N=1}^I y_i \left[\frac{1}{M} \sum_{j=1}^I A_j Z_{ij} \right]. \quad (6)$$

Puisque

$$\bar{A}_i = \sum_{M=1}^I A_j Z_{ij},$$

le terme entre crochets du côté droit de (6) est égal à 1, et

$$Y^* = \sum_{N=1}^I y_i = Y. \quad \text{C.Q.F.D.} \quad (7)$$

Par définition, la variance de Y est

$$V(Y) = \left(\frac{1}{M} \sum_{M=1}^I A_j \left(\frac{A_j}{A_j} - Y \right)^2 \right). \quad (8)$$

L'estimation de la variance de l'échantillon est alors (Cochran 1977):

$$v(Y) = \frac{1}{m(m-1)} \sum_{j=1}^m \left(\frac{A_j}{A_j} - Y \right)^2. \quad (9)$$

Le raisonnement général décrit par les équations (1) à (9) peut être appliqué à n'importe quel type particulier d'échantillonnage forestier conforme au processus en deux parties qui consiste à sélectionner des arbres à partir de points choisis de façon aléatoire.

À titre d'exemple additionnel de l'utilisation du concept du casse-tête, nous allons examiner la base d'échantillonnage obtenue quand l'échantillonnage par points est utilisé pour mesurer la croissance d'une forêt. Pour obtenir le maximum d'efficacité, on prend des mesures à deux moments différents, en utilisant les deux fois les mêmes points choisis au hasard. Ce genre d'échantillonnage relatif à la croissance des forêts est un échantillonnage par points avec deuxième mesure; ce type d'échantillonnage a fait l'objet d'un vaste traitement dans la littérature, le plus récemment par Van Deusen et coll. (1986) et Roesch et coll. (1989, 1991, 1993). Si un échantillonnage par points de ce type est effectué, et que la figure 1 représente le temps 1, le casse-tête pour l'échantillon global pourrait être découpé

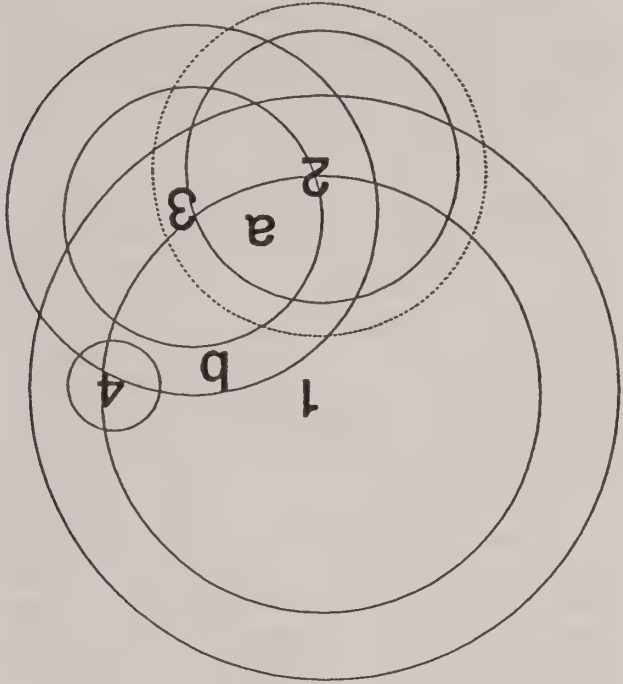


Figure 2. Les Pièces du casse-tête définies par emplacement, taille et temps. Un exemple d'unités d'échantillonnage dans le cas d'un échantillonnage par points avec deuxième mesure. Les arbres 1 et 3 ont poursuivi leur croissance et survécu, l'arbre 2 a continué de croître pendant quelque temps avant de mourir et l'arbre 4 est une recrue.

avec remplacement). Un estimateur non biaisé du total de la valeur d'intérêt pour un échantillon prélevé avec probabilité proportionnelle à la taille est donné par:

$$Y = \frac{A_T}{m} \sum_{j=1}^m \frac{A_j}{y_j} \qquad Y^* = \frac{A_T}{m} \sum_{j=1}^m \frac{A_j}{y_j^*} W_j, \tag{2}$$

où:

$$A_T = \sum_{j=1}^M A_j; \text{ la superficie totale de la forêt en acres,}$$

m = le nombre de points de l'échantillon,

M = le nombre de segments de terrain, et

W_j = le nombre de fois que la j ième unité apparaît dans l'échantillon.

Notons que W_j est un entier entre 0 et m inclusivement. A_j et y_j sont fixes et W_j est aléatoire. En outre, nous définissons:

$$Y = \sum_{i=1}^N y_i; \text{ le total de la valeur d'intérêt pour l'ensemble des arbres, et}$$

$$Y^* = \sum_{j=1}^M y_j; \text{ le total de la valeur d'intérêt pour l'ensemble des segments.}$$

Pour montrer que Y est non biaisé pour Y , nous allons d'abord montrer que Y est non biaisé pour Y^* , puis que Y^* est égal à Y . De la même façon que Cochran (1977, p. 252-255), nous pouvons montrer que Y est non biaisé pour Y^* :

$$E[Y] = E\left[A_T \sum_{j=1}^M \frac{A_j}{y_j} W_j\right]$$

$$= \frac{A_T}{m} \sum_{j=1}^M \frac{A_j}{y_j} E[W_j]. \tag{3}$$

W_j est une variable aléatoire multinomiale dont l'espérance est égale à $m(A_j/A_T)$. Par conséquent

$$E[Y] = \sum_{j=1}^M y_j = Y^*. \tag{4}$$

choisie, et s'il tombe dans le segment d , les trois arbres sont choisis, etc. L'arbre l serait donc sélectionné à partir des segments b , c , d et e . Cela donne une situation qui ressemble à celle décrite dans Kish (1965, section 11.2), c'est-à-dire que si les arbres formaient nos unités primaires d'échantillonnage, notre base d'échantillonnage serait une liste comprenant des inscriptions répétées de la même unité. Dans ce cas, la liste serait formée de grappes d'arbres, et la plupart des arbres seraient associés à plus d'une grappe. Les grappes sont sélectionnées avec probabilité proportionnelle à la taille du segment de terrain. La technique courante de pondération des éléments répétés d'une liste, examinée par Kish, se fonde plutôt sur la sélection d'unités primaires avec probabilité égale.

La méthode du casse-tête, en un sens, réduit la complexité du mécanisme d'échantillonnage, en appliquant d'abord la population d'arbres sur la population des segments de terrain, et en réduisant ainsi la base d'échantillonnage, formée d'une liste de grappes d'arbres dans laquelle les arbres appartiennent à plus d'une grappe, à une liste de segments de terrain uniques. Notre affirmation ci-dessous selon laquelle les simulations d'échantillonnage de forêts peuvent être simplifiées par la méthode du casse-tête est entièrement justifiée par la comparaison entre le coût initial de cette simplification de la liste d'échantillonnage et le besoin de prélever de nombreux échantillons à partir de cette liste.

Pour l'application de la population des arbres sur la population de segments, il faudrait préféablement qu'une observation, pour un segment, soit la somme de valeurs pondérées associées aux arbres, le poids pour chaque arbre étant proportionnel à sa probabilité d'être observé à partir de ce segment particulier. La probabilité qu'un arbre de l'échantillon i ait été sélectionné à partir du segment de terrain particulier j est:

$$p_{ij} = \left(\frac{A_j}{A_i}\right) Z_{ij},$$

où:

A_j = superficie du segment j en acres, et

$$Z_{ij} = \begin{cases} 1 & \text{si le segment } j \text{ fait partie du cercle } k \text{ de l'arbre } i \\ 0 & \text{sinon.} \end{cases}$$

La somme sur j des p_{ij} est 1. Nous pouvons maintenant écrire l'observation pour chaque segment sous forme d'une somme de valeurs pondérées associées aux arbres:

$$y_j = \sum_{i=1}^N p_{ij} y_i. \tag{1}$$

Supposons maintenant que nous choisissons au hasard m points à la surface d'une forêt, avec les mêmes hypothèses que ci-dessus (nous effectuons un échantillonnage

le point milieu est le centre de l'arbre, et dont le rayon est d dans le cas de l'échantillonnage par placettes et ar , dans le cas de l'échantillonnage par points. La zone de sélection de l'arbre i , de taille A_i (en acres), est la partie du cercle K de l'arbre i qui est à l'intérieur de la forêt, et c'est la zone pour laquelle l'arbre sera inclus dans l'échantillon si un point choisi au hasard se trouve dans cette zone.

Dans leur analyse de l'échantillonnage par points, Palley et Horwitz (1961) affirment que "... l'unité primaire d'échantillonnage est une grappe d'arbres associée à un lieu d'origine. Le lieu d'origine est un point dans le cas de l'échantillonnage par points ...". En fait, le lieu d'origine n'est pas un point, parce que la grappe d'arbres n'est pas sélectionnée uniquement à partir de ce point, mais plutôt à partir d'un ensemble infini de points situés à l'intérieur d'une zone précise.

Nous proposons l'approche différente qui consiste à considérer que les unités d'échantillonnage sont les sections mutuellement exclusives de terrain résultant de la superposition des zones de sélection des arbres individuels de la forêt.

La façon dont le terrain est découpé en unités primaires d'échantillonnage est clairement illustrée à la figure 1. La correspondance entre la population, la base d'échantillonnage et l'unité d'échantillonnage, comme elle est présentée par exemple dans Cochran (1977, p. 6), apparaît clairement: la population (ou l'image du casse-tête) est divisée en unités d'échantillonnage exhaustives et mutuellement exclusives (les pièces du casse-tête) qui, ensemble, forment la base d'échantillonnage. Chaque segment de terrain a une probabilité de sélection définie, et le total de ces probabilités pour l'ensemble des segments est 1. Nous allons appeler cette approche le concept du casse-tête.

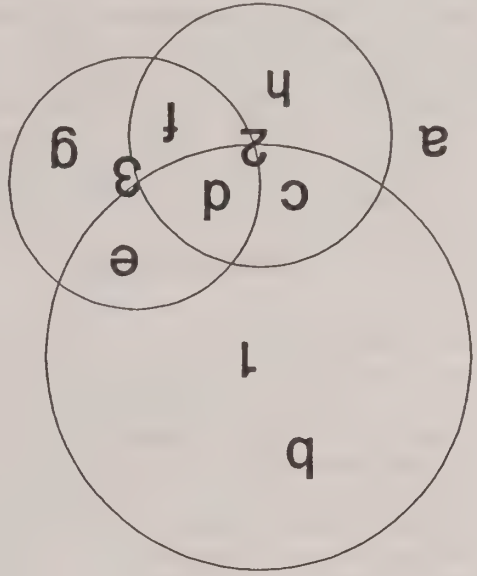


Figure 1. Les pièces du casse-tête. Les arbres 1, 2 et 3 ont leur centre là où se trouve le chiffre correspondant. Les cercles qui les entourent représentent les zones de sélection des arbres. Chaque segment désigné par une lettre constitue une unité d'échantillonnage.

À chaque segment de terrain sont associés des attributs les attributs.

L'élément essentiel est que chaque point est équivalent aux autres à l'intérieur d'un segment particulier. Les segments de terrain sont évidemment sélectionnés avec probabilité proportionnelle à la taille. Dans le cas de l'échantillonnage par points, la taille des segments est déterminée par les surfaces terrestres et la distribution spatiale des arbres, ainsi que par la constante α choisie. Dès que α est choisie, la base d'échantillonnage à un point du temps donné est fixée. Dans le cas de l'échantillonnage par placettes, la taille du segment est déterminée par d et par la distribution spatiale des arbres. Par conséquent, peu importe la méthode utilisée pour déterminer les arbres à inclure dans l'échantillon (p. ex. échantillonnage par placettes ou échantillonnage par points), on peut imaginer que dans chaque cas, on découpe l'image du casse-tête d'une certaine façon, on sélectionne des pièces avec probabilité proportionnelle à la taille et on retourne chaque pièce pour en lire à l'endos les attributs.

À l'appui de notre affirmation énoncée plus haut selon laquelle cette méthode est souvent celle qui convient le mieux, notons que l'objectif de la plupart des échantillonnages forestiers est de décrire la forêt, et non les arbres individuels. Nos agrégations sont souvent faites par acre ou par hectare, c.-à-d. des unités de la terre forestière, et non des unités de l'arbre. À partir du même endroit, nous pouvons faire des mesures portant sur de nombreux éléments autres que les arbres, par exemple les caractéristiques topographiques et d'autres données sur le site. Il est donc préférable, en général, de considérer des parties de la forêt, plutôt que des arbres individuels, comme les unités d'échantillonnage.

Nous allons nous limiter, dans la présente communication, à un traitement général de l'échantillonnage des forêts, mais notre analyse s'applique à n'importe quel type particulier d'échantillonnage des forêts qui se fonde sur la sélection d'arbres selon une fonction liée à des points choisis au hasard. La seule différence est la définition des segments de terrain, c'est-à-dire la façon dont l'image du casse-tête est découpée. Par exemple, dans l'échantillonnage par placettes, le terrain est divisé en pièces définies par des cercles superposés de taille égale, tandis que dans l'échantillonnage par points, la définition fait intervenir des cercles superposés de tailles proportionnelles à la surface terrestre de chaque arbre.

Imaginons, pour les besoins de notre analyse, qu'un point est choisi au hasard à la surface d'une forêt et qu'une fonction quelconque est utilisée pour sélectionner les arbres qui feront partie de l'échantillon. Supposons aussi que notre forêt compte trois arbres (1, 2 et 3) dont les zones de sélection se chevauchent. À la figure 1, le centre des arbres 1, 2 et 3 se trouve là où le chiffre apparaît, et des cercles délimitent la zone de sélection de ces arbres. Chaque segment désigné par une lettre représente une unité d'échantillonnage différente. Si le point tombe dans le segment a , la grappe vide est choisie, s'il tombe dans le segment b , la grappe contenant seulement l'arbre 1 est

Un nouveau concept pour l'échantillonnage des forêts

FRANCIS A. ROESCH, JR., EDWIN J. GREEN et CHARLES T. SCOTT¹

RÉSUMÉ

Un concept généralisé est présenté pour toutes les méthodes couramment utilisées d'échantillonnage des forêts. Selon ce concept, la forêt est perçue comme une image bidimensionnelle découpée en pièces comme un casse-tête, les pièces étant définies par les probabilités de sélection individuelles des arbres de la forêt. Ce concept produit un nombre fini d'unités d'échantillonnage sélectionnées de façon indépendante, contrairement à tous les autres concepts généralisés d'échantillonnage des forêts présentés jusqu'à maintenant.

MOTS CLÉS: Échantillonnage des forêts; échantillonnage ppt.

1. INTRODUCTION

L'échantillonnage des forêts est souvent accompli sous forme d'un processus en deux parties: un point de la forêt est d'abord choisi au hasard, puis une grappe d'arbres entourant ce point est incluse dans l'échantillon en vertu d'une règle donnée. Les deux règles les plus courantes sont l'échantillonnage par placettes (circulaires, à superficie fixe) et l'échantillonnage par points (horizontal). Dans le premier cas, tous les arbres pour lesquels le centre de la section transversale du tronc à 4.5 pieds au-dessus du sol se situe à l'intérieur d'une distance horizontale constante (d) du point choisi au hasard sont inclus dans l'échantillon. Dans le second cas, l'arbre i est ajouté à l'échantillon si ce centre se trouve à l'intérieur d'une distance horizontale αr_i du point choisi au hasard, où r_i est le rayon de la section transversale et α est une constante, choisie en fonction de l'intensité d'échantillonnage désirée. L'arbre i est sélectionné avec une probabilité proportionnelle à πd^2 dans l'échantillonnage par placettes (la probabilité est la même pour tous les arbres) et avec une probabilité proportionnelle à πr_i^2 (surface terrière de l'arbre i) dans l'échantillonnage par points (les plus gros arbres ont une plus grande probabilité d'être choisis).

L'élément constituant l'unité d'échantillonnage, dans les diverses méthodes d'échantillonnage des forêts, est une question dont ont abondamment débattu les auteurs du domaine de la foresterie. Certains considèrent l'arbre comme l'unité d'échantillonnage (p. ex. Oderwald 1981), tandis que pour d'autres, la grappe d'arbres associée au point (p. ex. Bailey et Horwitz 1961; Schröder 1970), la placette circulaire (p. ex. Cunia 1965), ou le point (p. ex. Husch 1955) constitue l'unité d'échantillonnage. Ces différents points de vue s'appuient sur divers outils statistiques. Par exemple, l'utilisation de l'arbre comme unité d'échantillonnage oblige à utiliser la théorie de l'échantillonnage de populations finies, tandis que l'utilisation du point comme unité d'échantillonnage exige le

recours à la théorie un peu plus avancée de l'échantillonnage de populations infinies. En outre, l'échantillonnage par placettes a été présenté, traditionnellement, comme une méthode ayant la placette comme unité d'échantillonnage, tandis que pour l'échantillonnage par points, l'unité d'échantillonnage utilisée a généralement été l'arbre ou le point. Par conséquent, la disparité apparente entre ces mécanismes d'échantillonnage très communs et très semblables est d'origine artificielle.

Nous allons présenter un concept d'unité primaire d'échantillonnage qui s'applique à tous les types de plans d'échantillonnage forestier en vertu desquels les arbres sont sélectionnés d'après l'emplacement d'un point choisi au hasard. Nous allons aussi montrer que ce concept est simple et qu'il donne un nombre fini d'unités d'échantillonnages mutuellement exclusives et prélevées de façon indépendante. Notons, par contraste, qu'à titre d'unités d'échantillonnage, ni les arbres ni les grappes ne possèdent à la fois ces deux caractéristiques; en effet, les arbres ne sont pas sélectionnés de façon indépendante et les grappes d'arbres ne sont pas mutuellement exclusives. Le concept que nous allons présenter diffère aussi de ceux où le point placé au hasard, ou encore la placette, est l'unité d'échantillonnage, car il y a dans ces derniers cas un nombre infini d'unités. Nous allons aussi faire valoir que ce concept, dans bien des cas, est celui qui convient le mieux aux besoins.

2. LE CONCEPT DU CASSE-TÊTE

Supposons que la forêt compte N arbres, et que ceux-ci sont désignés 1, 2, ..., N . À ces N arbres sont associées les valeurs d'intérêt $y = \{y_1, y_2, \dots, y_N\}$, des cercles K dénotés $K = \{K_1, K_2, \dots, K_N\}$ et des zones de sélection ayant les tailles $A = \{A_1, A_2, \dots, A_N\}$. Crosebaugh et Stover (1957) ont d'abord défini les cercles K dans le contexte de l'échantillonnage par points. Pour nos besoins, le cercle K de l'arbre i , K_i , est un cercle imaginaire, dont

¹ Francis A. Roesch, Jr., Mathematical Statistician, Institute for Quantitative Studies, Southern Forest Experiment Station, USDA Forest Service, 701 Loyola Avenue, New Orleans, LA 70113; Edwin J. Green, Professor of Forestry, Cook College-Rutgers University, P.O. Box 231, New Brunswick, NJ 08903; et Charles T. Scott, Project Leader, Forest Ecosystem Modeling Unit, Northeastern Forest Experiment Station, USDA Forest Service, 359 Main Road, Delaware, Ohio 43015.

On constate que dans la plupart des cas, l'échantillon-
nage à partir d'une base imparfaite est plus efficient.

REMERCIEMENTS

Les auteurs tiennent à remercier un des rédacteurs asso-
ciés et les arbitres pour leurs conseils utiles qui ont contribué
à améliorer cet article.

BIBLIOGRAPHIE

HANSEN, M.H., HURWITZ, W.N., et JABINE, T.N. (1963).
The Use of imperfect lists for probability sampling at the U.S.
Bureau of Census. *Bulletin de l'Institut Internationale de
Statistique*, 40, 497-517, (avec discussion).

INDIAN STATISTICAL INSTITUTE (1988). *A study of
Fishermen in West Bengal: 1985-1986.*

RAO, J.N.K. (1968). Some non-response sampling theory when
the frame contains an unknown amount of duplication.

Journal of the American Statistical Association, 63, 87-90.
SINGH, R. (1977). A note on the use of incomplete multi-
auxiliary information in sample surveys. *Australian Journal
of Statistics*, 19, 105-107.

SINGH, R. (1983). On the use of incomplete frames in sample
surveys. *Biometrical Journal*, 25, 545-549.

SZAMEITAT, K., et SCHAFFER, K.A. (1963). Imperfect frames
in statistics and the consequences for their use in sampling.
Bulletin de l'Institut Internationale de Statistique, 40, 517-538,
(avec discussion).

WRIGHT, T., et TSAO, H.J. (1983). *A frame on frames: An
annotated bibliography. Statistical Methods and Improvement
of Data Quality*, (Ed. T. Wright). New York: Academic
Press, 25-72.

où W défini en (2.3).

Pour évaluer l'effet d'un échantillonnage dans la liste
des bénéficiaires, qui sont au nombre de 27, nous calcu-
lons des estimations des variables suivantes:

R = frais de régénération à l'acre;

X = superficie moyenne des plans d'eau exploités, par

entreprise (en acres);

NX = superficie totale des plans d'eau exploités par

l'ensemble des 23 entreprises.

Le tableau ci-dessous donne l'efficacité (ρ) pour

diverses valeurs de m et n .

Tableau 1

Efficiéce de l'échantillonnage dans une base parfaite par
rapport à l'échantillonnage dans une base imparfaite (ρ)

Taille d'échantillon		Efficiéce pour les estimateurs de		
n	m	R	X	NX
2	2	0.8695	0.6453	0.9508
4	4	0.8841	0.6561	0.9668
6	6	0.9022	0.6696	0.9866
8	8	0.9225	0.6866	1.0117
8	9	0.7791	0.5781	0.8519
10	10	0.9551	0.7088	1.0444
10	11	0.8172	0.6065	0.8937

relativement plus élevées se répètent dans la base imparfaite. Comme σ_y^2 est fixe pour un ensemble donné de N valeurs W , il peut y avoir des cas où ρ , tel qu'il est défini en (3.4), est plus petit que 1 (de fait, $S(2, W)$ est égal à No_2^2 lorsque les valeurs W sont toutes égales à 0) et par conséquent, il y aura des cas où l'échantillonnage à partir d'une base imparfaite sera préférable à l'échantillonnage à partir d'une base complète, même si la fraction de sondage est moindre dans la première situation.

3.2 Estimation d'une moyenne

Comme nous l'avons vu dans la section 3.1, y^* est un estimateur non biaisé de $(NY)/M$, où M est connu mais N , inconnu. Il est donc nécessaire d'estimer N pour obtenir un estimateur de Y . Notons que c^* est un estimateur non biaisé de (N/M) et que, par conséquent,

$$\hat{Y} = y^*/c^*$$

est un estimateur naturel de Y du type "estimateur par quotient". Si on remplace c^* , défini dans la section 3.1, par c^* , l'EQM de \hat{Y} est rendue par l'expression

$$EQM(\hat{Y}) = \frac{M(M - m)}{mN^2(M - 1)} \{No_2^2 - S(2, D)\},$$

où les valeurs D sont définies en (2.3). Si on remplace W , dans la section 3.1, par D , on peut conclure que (3.6) se vérifie et qu'il est plus avantageux d'utiliser une base imparfaite lorsque (2.5) est vraie.

3.3 Estimation d'un total

Si on veut estimer un total, par exemple NY , en fonction d'un EASSR de taille m tiré d'une base imparfaite, on utilise habituellement l'estimateur

$$(\widehat{NY}) = MY^*,$$

qui est non biaisé pour NY et dont la variance est définie

$$EQM(MY^*) = \text{Var}(MY^*)$$

$$= \frac{m(M - m)}{m(M - 1)}$$

$$\left\{ No_2^2 - S(2, Y) + (NY)^2 \left(\frac{1}{N} - \frac{1}{M} \right) \right\}.$$

On peut exprimer ρ par la formule

$$\rho = \frac{nM(M - m)(N - 1)}{mN(N - n)(M - 1)}$$

$$\left\{ 1 - \frac{S(2, Y) - (NY)^2(1/N - 1/M)}{No_2^2} \right\}.$$

$$\left\{ S(2, Y) - (NY)^2 \left(\frac{1}{N} - \frac{1}{M} \right) \right\} / No_2^2, \quad (3.7)$$

Il est clair d'après l'expression de $\text{Var}(MY^*)$ que

$$\sum_{i=1}^N (Y_i - Y)^2 \geq \sum_{j=1}^M (Y_j^* - Y^*)^2, \quad (3.8)$$

l'expression (3.7) sera toujours non négative. On peut alors tirer des conclusions semblables à celles énoncées dans la section 3.1, par exemple, (3.6) se vérifie lorsque (2.5) est vraie.

4. EXEMPLE

Comme nous l'avons souligné plus haut, les unités d'échantillonnage qui devaient servir à l'enquête sur les pêcheurs étaient tirées de la liste de pêcheurs-bénéficiaires qui existait à ce moment-là. Puisque cette enquête était polyvalente, elle a permis d'observer de nombreuses caractéristiques des unités d'échantillonnage qui avaient rapport soit au ménage, soit à l'entreprise de pêche à laquelle appartenait l'unité d'échantillonnage. Comme on ne connaît pas le nombre de bénéficiaires (M), le nombre d'entreprises ou de ménages correspondants (N) étant inconnu, il n'a pas été possible d'évaluer l'effet de l'utilisation d'une base imparfaite pour cette enquête. Cependant, pour les besoins de cet article, nous allons considérer comme une population en soi les échantillons qui ont été tirés dans une région géographique (portion d'un district administratif du Bengale-Occidental) et nous allons étudier l'effet d'un rééchantillonnage dans cette population. On dénombre dans cette région 27 bénéficiaires (M) et 23 entreprises distinctes (N); dix-neuf de ces entreprises ont un propriétaire unique (N_1) et les quatre autres sont des entreprises à propriété conjointe (N_2). Les caractéristiques qui nous intéressent sont le coût de la régénération des plans d'eau (Y) et la superficie des plans d'eau exploités (X).

Les statistiques sommaires de Y et de X sont les suivantes:

$$\sum Y_i = 58,815, \quad \sum X_i = 23,36,$$

$$R = \left(\sum Y_i \right) / \left(\sum X_i \right) = 2,517,77,$$

parfaite par rapport à une base imparfaite, pour n importe quel estimateur, par l'expression

$$\rho = \frac{\text{EQM d'un échantillon de taille } m \text{ provenant de la base imparfaite}}{\text{EQM d'un échantillon de taille } n \text{ qui proviendrait d'une base parfaite}} \quad (2.4)$$

Posons aussi f comme la fraction de sondage commune lorsque les fractions de sondage sont les mêmes, c.-à-d.,

$$n = fN, \quad m = fM = n(1 + \alpha). \quad (2.5)$$

3. RÉSULTATS

Avant de répondre à la question fondamentale que nous nous sommes posée dans la section 1, à savoir s'il faut échantillonner à partir d'une base parfaite ou d'une base imparfaite, nous allons considérer brièvement les deux options dans une perspective de coûts. Si on prévoit que la mise à jour de la base imparfaite coûtera plus cher que la collecte de données qui devra se faire auprès des $(m - n)$ unités additionnelles, il est plus avantageux d'utiliser la base imparfaite avec un échantillon plus grand que de mettre à jour cette base; il en est ainsi lorsque

$$\frac{b_1}{b_0} \left(\frac{m}{m - n} - \frac{N}{n} \right) \leq 1, \quad (3.1)$$

où b_1 est le coût unitaire de la collecte de données et b_0 , le coût unitaire de la mise à jour. Notons qu'il faut visiter effectivement N unités pour mettre à jour la base incomplète puisque les $(M - N)$ autres unités sont des répétitions et qu'elles peuvent être identifiées grâce à l'hypothèse b). Notons aussi que même dans le cas d'un EASSR issu de la base imparfaite, le nombre supplémentaire d'unités à visiter est tout au plus de $(m - n)$ puisque l'échantillon peut compter plusieurs répétitions de la même unité sous des identités différentes. Ces observations conduisent à l'inégalité (3.1), ce qui confirme la préférence pour une base imparfaite.

Comme nous l'avons mentionné dans la section 1, la base imparfaite ne permettra pas de connaître le nombre total d'unités de population, N . L'estimation d'une moyenne et l'estimation d'un total sont donc des problèmes différents; le problème de l'estimation du ratio, mais essentiellement le problème de l'estimation d'estimer directement et sans biais un total à partir d'un EASSR de taille m tiré de la base imparfaite. Par conséquent, il est convainable d'estimer un ratio de population (procédé analogue à l'estimation pour domaine), et de considérer l'estimation de moyenne comme un cas particulier de l'estimation de ratio, puis d'envisager séparément l'estimation d'un total.

3.1 Estimation d'un ratio

Pour l'estimation du ratio $R = (Y/X)$, l'estimateur habituel est

$$R = \bar{y}^*/\bar{x}^*,$$

où les lettres minuscules représentent les quantités correspondantes pour un échantillon; \bar{y}^* est la moyenne des valeurs Y^* établie d'après un échantillon de taille m issu de la base imparfaite, etc.; \bar{y}^* et \bar{x}^* sont des estimateurs non biaisés de $(N\bar{Y}/M)$ et de $(N\bar{X}/M)$ respectivement. En utilisant la méthode delta, on peut définir une formule approximative pour l'EQM de R , $E(R - R)^2$, à savoir

$$\frac{M - m}{M} \sum_{i=1}^t M \frac{m(X_i^*)^2 (M - 1)}{W_i^2}; \quad (3.2)$$

utilisant les relations de la section 2, on peut réécrire (3.2) sous la forme

$$\text{EQM}(R) = \frac{M(M - m)}{m(N\bar{X})^2 (M - 1)} \{N\sigma_w^2 - S(2, W)\},$$

où les valeurs W sont définies en (2.3) et les valeurs W^* calculées de la même manière. Une conséquence de (2.2) est que $S(2, W) \geq 0$, ce qui nous permet d'écrire, d'après (3.2),

$$0 \leq 1 - \frac{N\sigma_w^2}{S(2, W)} \leq 1. \quad (3.3)$$

Maintenant d'après l'équation (2.4), nous pouvons exprimer l'efficacité ρ par la formule

$$\rho = \frac{nM(M - m)(N - 1)}{mN(N - n)(M - 1)} \left\{ 1 - \frac{N\sigma_w^2}{S(2, W)} \right\}. \quad (3.4)$$

Lorsque les fractions de sondage sont égales, ρ peut être exprimée comme suit:

$$\rho = \frac{(1 + \alpha)(N - 1)}{(1 + \alpha)(N - 1) + \alpha} \left\{ 1 - \frac{N\sigma_w^2}{S(2, W)} \right\}. \quad (3.5)$$

On peut donc affirmer, si l'on s'appuie sur (3.3), que ρ , telle qu'elle est définie en (3.5), satisfait l'inégalité

$$0 \leq \rho \leq 1 \quad (3.6)$$

et que, par conséquent, il est avantageux d'utiliser une base imparfaite pour estimer un ratio.

Il convient de souligner que $S(2, W)$ est non décroissant en α et que pour une valeur α fixe, $S(2, W)$ a une valeur plus élevée lorsque les unités qui ont des valeurs W

Dans les deux enquêtes sur les ménages de pêcheurs, on pensait que la plupart des variables économiques étudiées étaient étroitement liées au nombre de membres de SCP ou de bénéficiaires de la FFDa au sein d'un ménage, en ce sens que la variabilité de ces caractéristiques par membre de SCP ou par bénéficiaire de la FFDa était inférieure à la variabilité des mêmes caractéristiques par ménage. On estimait donc que l'utilisation d'une base imparfaite était acceptable dans ces conditions.

Nous verrons que dans des cas comme ci-dessus, l'estimateur d'un ratio, d'une moyenne ou d'un total peut avoir une erreur quadratique moyenne moins élevée même si la fraction de sondage est plus faible dans la base imparfaite que celle utiliser dans la base parfaite.

Même si la variabilité n'a pas de rapport avec le nombre d'enregistrements répétés, nous verrons que pour l'estimation d'un ratio ou d'une moyenne, la base imparfaite doit être préférée à la base parfaite, du point de vue de l'EQM, lorsque les fractions de sondage des deux bases sont les mêmes.

2. NOTATION ET RELATIONS

Considérons une population finie constituée de N unités U_1, U_2, \dots, U_N . Désignons par $U_1^*, U_2^*, \dots, U_M^*$ les unités listées dans une base imparfaite. Pour $k = 1, 2, \dots, r$, désignons par A_k une sous-population des N unités formée de N_k unités distinctes. Chacune des unités de A_k est listée exactement k fois dans la base imparfaite sous des identités différentes. Supposons que

a) chaque U_i appartient à une sous-population A_k , pour une valeur k quelconque (c'est-à-dire que chaque U_i est incluse dans la base imparfaite au moins une fois) et que

b) si U_j^* est échantillonnée à l'aide de la base imparfaite, il sera possible de déterminer, à l'étape de la collecte de données, l'unité U_i correspondante et la valeur de k (c.-à-d. le nombre de fois que U_i figure dans la base incomplète sous différentes identités, l'une d'elles correspondant à l'unité échantillonnée U_j^*) pour laquelle U_i appartient à A_k .

Les relations suivantes sont valides:

$$N_1 + N_2 + \dots + N_r = N;$$

$$N_k \geq 0, k = 1, 2, \dots, r,$$

$$N_1 + 2N_2 + \dots + rN_r = M,$$

où r, N_1, N_2, \dots, N_r et N sont tous inconnus, seul M étant connu $M \geq N$; M peut s'écrire, pour une valeur α inconnue,

$$M = N(1 + \alpha), \quad \alpha \geq 0. \quad (2.1)$$

Désignons les valeurs X et Y pour l'unité U_i par X_i et Y_i respectivement, ($i = 1, 2, \dots, N$). Puisque chaque U_j^* ($j = 1, 2, \dots, M$), peut être identifiée à une U_i pour un i quelconque, ($i = 1, 2, \dots, N$) et puisque U_i appartient à A_k , pour une valeur k quelconque, ($k = 1, 2, \dots, r$), définissons les valeurs X, Y et C pour l'unité U_j^* par les équations suivantes:

$$X_j^* = X_i/k, \quad Y_j^* = Y_i/k, \quad C_j^* = 1/k.$$

En raison des hypothèses a) et b), les valeurs X^*, Y^* et C^* sont observables pour les unités échantillonnées qui proviennent de la base imparfaite.

Les identités suivantes mettent en relation les paramètres tirés de la base imparfaite et celles tirées de la base parfaite:

$$\sum_{j=1}^M Y_j^* = MY^* = \sum_{i=1}^N Y_i = NY, \\ \sum_{j=1}^M C_j^* = MC^* = N;$$

$$\sum_{j=1}^M (Y_j^* - \bar{Y})^2 = N\sigma_Y^2 - S(2, Y)$$

$$+ (N\bar{Y})^2(1/N - 1/M),$$

où

$$N\sigma_Y^2 = \sum_{i=1}^N (Y_i - \bar{Y})^2$$

et

$$S(a, Z) = \sum_{i=1}^k (1 - 1/k) \left\{ \sum_{i: U_i \in A_k} Z_i^a \right\}; \quad (2.2) \\ \sum_{j=1}^M (C_j^* - \bar{C})^2 = N(1 - N/M) - S(0, Y); \\ \sum_{j=1}^M (Y_j^* - \bar{Y})(C_j^* - \bar{C})$$

Pour l'unité U_i , posons

$$D_i = Y_i - \bar{Y}, \quad W_i = Y_i - R X_i, \quad \text{où } R = \bar{Y}/\bar{X}.$$

(2.3)

Comme on ne suppose pas l'existence d'information supplémentaire sur les unités, les comparaisons se feront en fonction d'un EASSR. Soit m la taille de l'échantillon tiré de la base imparfaite et n , l'effectif correspondant s'il s'agit de la base parfaite. Définissons l'efficacité d'une base

Echantillonnage dans des bases imparfaites contenant un nombre inconnu d'enregistrements répétés

SHIBDAS BANDYOPADHYAY et A.K. ADHIKARI¹

RÉSUMÉ

Dans cette étude, nous nous intéressons à des bases de sondage imparfaites desquelles on n'a retiré aucune unité de population mais dans lesquelles un nombre indéterminé d'unités peuvent avoir été ajoutées un nombre indéterminé de fois sous des identités différentes. Lorsqu'on ne pose pas l'hypothèse de l'existence d'information supplémentaire concernant des unités de la base imparfaite, il est établi qu'en ce qui a trait à l'estimation d'un ratio ou d'une moyenne de population, l'erreur quadratique moyenne des estimateurs fondés sur la base imparfaite est inférieure à celle des estimateurs fondés sur la base parfaite pour l'échantillonnage aléatoire simple, lorsque les fractions de sondage des deux bases sont les mêmes. Cependant, cette relation n'est pas toujours vraie en ce qui concerne l'estimation d'un total de population. Il peut aussi arriver que l'estimateur d'un ratio, d'une moyenne ou d'un total ait une erreur quadratique moyenne moins élevée même si la fraction de sondage est plus faible dans la base imparfaite que celle utilisée dans la base parfaite.

MOTS CLÉS: Base de sondage imparfaite; efficacité.

1. INTRODUCTION

L'absence de bases de sondage complètes est un problème fréquent dans la planification d'enquêtes. L'Institut international de statistique a reconnu l'importance de s'intéresser au problème de l'échantillonnage dans des bases imparfaites et a, de fait, inscrit cette question au programme de sa 34^{ème} session, tenue à Ottawa (Canada). À cette occasion, Hansen et coll. (1963) et Szameitai et Schaffer (1963) avaient présenté des communications. Nous renvoyons aussi le lecteur aux ouvrages de Singh (1977, 1983). Wright et Tsoa (1983) ont rédigé une bibliographie sur les bases de sondage afin de mettre en lumière les problèmes qui se posent lorsqu'on fait des sondages à l'aide de bases imparfaites.

Ces dernières années, l'Indian Statistical Institute a réalisé deux enquêtes indépendantes pour évaluer l'effet de programmes gouvernementaux sur la condition économique de la communauté de pêcheurs du Bengale-Occidental, en Inde. Dans la première enquête (1988), les ménages étaient choisis à même les listes de membres des sociétés coopératives de pêcheurs (SCP). Dans la seconde enquête, plus récente, on s'est servi de la liste des pêcheurs qui recevaient de l'aide de la Fish Farmer's Development Agency (FFDA). On savait que les membres des SCP ou les bénéficiaires de la FFDA n'appartenaient pas tous à des ménages différents, mais il n'était pas possible de déterminer quels membres ou quels bénéficiaires appartenaient au même ménage sans communiquer avec les ménages. Par conséquent, lorsqu'on s'est servi des listes de membres des SCP ou des listes de bénéficiaires de la FFDA pour le choix des ménages, les bases de sondage contenaient un nombre inconnu de doubles enregistrements. Comme les données sur le ménage étaient recueillies au moyen d'une interview sur place, on pouvait détecter les

doubles enregistrements uniquement parmi les ménages échantillonnés. On a divisé les valeurs des variables observées pour les ménages de l'échantillon par le nombre correspondant d'enregistrements répétés présents dans la base de sondage tandis qu'on a conservé dans l'échantillon, sous leurs diverses identités, les ménages qui se répétaient. L'exemple de bases imparfaites que nous étudions ici est un cas particulier de Rao (1968). Un des arbitres a souligné que la situation que nous décrivons dans cet article existe à Statistique Canada pour certaines bases utilisées dans les enquêtes-entreprises.

Le genre de bases imparfaites que nous étudions ici sont celles desquelles on n'a retiré aucune unité de population mais dans lesquelles un nombre indéterminé d'unités peuvent avoir été ajoutées un nombre indéterminé de fois sous des identités différentes. On suppose qu'il est possible d'évaluer, à l'étape de la collecte des données, le nombre d'enregistrements répétés présents dans la base de sondage pour chaque unité échantillonnée. On n'exclut pas la possibilité de tirer plusieurs enregistrements de la même unité de population. On ne suppose pas l'existence d'information supplémentaire concernant les unités de la base imparfaite et on s'en tient uniquement aux plans d'échantillonnage aléatoire simple sans remise (EASSR).

Comme les bases imparfaites que nous étudions ici ne nous permettront pas de connaître le nombre total d'unités de population, l'estimation de la moyenne et l'estimation du total pour un caractère de la population ne posent pas les mêmes problèmes.

La question fondamentale à laquelle nous allons tenter de répondre dans cet article est la suivante. Quelle est la meilleure voie à suivre: mettre à jour la base de sondage imparfaite et tirer un échantillon, ou utiliser la base imparfaite telle quelle?

¹ Shibdas Bandyopadhyay et A.K. Adhikari, Indian Statistical Institute, Calcutta, Inde 700 035.

- HANSEN, M.H., MADOW, W.G., et TEPPING, B.J. (1983). An evaluation of model-dependent and probability-sampling inferences in sample surveys. *Journal of the American Statistical Association*, 78, 776-796.
- HIDIROGLOU, M., et SÄRNDALE, C.-E. (1989). Small domain estimation: a conditional analysis. *Journal of the American Statistical Association*, 84, 266-275.
- HOLT, D., et SMITH, T.M.F. (1979). Post stratification. *Journal of the Royal Statistical Society A*, 142, 33-46.
- KIEFER, J. (1977). Conditional confidence statements and confidence estimators (avec discussion). *Journal of the American Statistical Association*, 72, 789-827.
- KREWSKI, D., et RAO, J.N.K. (1981). Inference from stratified samples: Properties of the linearization, jackknife, and balanced repeated replication methods. *Annals of Statistics*, 9, 1010-1019.
- LITTLE, R.J.A. (1991). Post-Stratification: A modeler's perspective. *Proceeding of the Section on Survey Research Methods, American Statistical Association*, à paraître.
- RAO, J.N.K. (1985). Inférence conditionnelle dans les enquêtes par sondage. *Techniques d'enquête*, 11, 17-35.
- RAO, J.N.K. (1992). Estimating Totals and Distribution Functions Using Auxiliary Information at the Estimation Stage. Présenté au Workshop on Uses of Auxiliary Information in Surveys, Statistics Sweden.
- RAO, J.N.K., et WU, C.F.J. (1985). Inference from stratified samples: Second order analysis of three methods for nonlinear statistics. *Journal of the American Statistical Association*, 80, 620-630.
- ROBINSON, J. (1987). Conditioning ratio estimates under simple random sampling. *Journal of the American Statistical Association*, 82, 826-831.
- ROYAL, R.M. (1971). Linear regression models in finite population sampling theory. Dans *Foundations of Statistical Inference*, (Eds. V.P. Godambe et D.A. Sprott). Toronto: Holt, Rinehart et Winston.
- SÄRNDALE, C.-E., SWENSSON, B., et WRETSMAN, J. (1989). The weighted residual technique for estimating the variance of the finite population total. *Biometrika*, 76, 527-537.
- SÄRNDALE, C.-E., SWENSSON, B., et WRETSMAN, J. (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.
- VALLIANT, R. (1990). Comparisons of variance estimators in stratified random and systematic sampling. *Journal of Official Statistics*, 6, 115-131.
- VALLIANT, R. (1993). Post-stratification and conditional variance estimation. *Journal of the American Statistical Association*, 88, 89-96.
- VATES, F. (1960). *Sampling Methods for Censuses and Surveys*, (3^{ème} Ed.). London: Griffin.

Cas 2. Dans cette situation, on peut vérifier que l'estimateur

$$B_2 = (1 - p') \left[\hat{Y} - \Sigma_{12} \Sigma_{22}^{-1} \left(\frac{\hat{Y}}{N_{..}} - M_2 \right) \right],$$

est, de façon approximative, conditionnellement non biaisé pour $\mu - \mu'$ et que, comme \hat{Y}_{LR} est conditionnellement non biaisé pour μ , il s'ensuit directement que l'estimateur

$$\hat{Y}_2 = \hat{Y}_{LR} - B_2 = p' \left[\hat{Y} - \Sigma_{12} \Sigma_{22}^{-1} \left(\frac{\hat{Y}}{N_{..}} - M_2 \right) \right].$$

est approximativement non biaisé, de façon tant conditionnelle que non conditionnelle, pour μ . On peut également vérifier que:

$$\text{var} [\hat{Y}_2 | N] = \text{var} [\hat{Y}_2] = m^{-1} [p' V_c p].$$

Outre les problèmes de l'estimateur de régression linéaire mentionnés plus haut, notons que cet estimateur, en général, n'est même pas bien défini, car les situations où les paramètres de la base $\{\phi_k, 1 \leq k \leq K\}$ sont connus quand la base est déficiente sont rares, voire inexistantes.

5. CONCLUSION

La présente étude a généralisé les techniques asymptotiques suggérées par Robinson (1987) pour étudier le problème de la stratification *a posteriori* selon une approche d'inférence conditionnelle, fondée sur le plan. Un article important sur l'étude conditionnelle de la stratification *a posteriori* a été publié par Holt et Smith (1979); dans cet article, l'une des hypothèses de base était que Y_{ps} est conditionnellement non biaisé. Ce sera vrai (du moins asymptotiquement) seulement si $I'(H - D(\mu_k)) = 0$; par conséquent, de façon générale, cette hypothèse est fausse. En fait, l'échantillonnage aléatoire simple d'unités élémentaires pourrait être l'un des rares cas où, en situation réelle, cette hypothèse de base est vérifiée.

Selon notre analyse conditionnelle, l'estimateur de régression linéaire est le meilleur choix parmi les quatre étudiés. C'est le seul qui soit conditionnellement non biaisé. L'estimateur de stratification *a posteriori* n'est pas meilleur (ni pire) que l'estimateur de Horvitz-Thompson ou l'estimateur par quotient; tous ont des termes de biais conditionnel d'ordre $m^{-1/2}$. Tous les estimateurs ont la même variance conditionnelle jusqu'aux termes d'ordre m^{-1} ; en outre, la variance conditionnelle ne dépend pas de N , le vecteur des proportions estimées dans les strates *a posteriori*. Par conséquent, du fait qu'il est conditionnellement non biaisé, l'estimateur de régression présente la plus faible erreur quadratique moyenne conditionnelle. Les estimateurs de Horvitz-Thompson, par quotient et de stratification *a posteriori* sont non biaisés de façon non conditionnelle. Bien que cela soit quelque peu illogique, on pourrait tenter de défendre ces estimateurs en comparant

La question de l'estimation de la variance est un aspect que nous n'avons pas examiné. Un estimateur de la variance fondé sur le plan pour l'estimateur de régression peut être obtenu à l'aide des méthodes de Särndal, Swensson et Wretman (1989).

Le problème posé par une base de sondage déficiente amène des complications qui n'existeraient pas autrement. Chacun des estimateurs de la moyenne étudiés ici est biaisé, tant de façon conditionnelle que de façon non conditionnelle. Des rajustements tenant compte de ce biais ne sont possibles qu'en vertu de l'hypothèse contraignante selon laquelle la moyenne des unités de chaque strate *a posteriori* est la même pour toutes les unités de la population, que celles-ci soient incluses dans la base ou qu'elles en soient exclues.

Leur propriétés non conditionnelles aux propriétés conditionnelles de l'estimateur de régression linéaire. Mais même dans cette perspective mixte, l'estimateur \hat{Y}_{LR} (théor) est nettement supérieur aux autres. Non seulement il est conditionnellement non biaisé, mais la variance conditionnelle de l'estimateur de régression linéaire ne peut être supérieure à la variance non conditionnelle de n'importe quel des autres estimateurs. Pour les grands échantillons d'UPD, la version empirique de l'estimateur de régression héritera de ces propriétés intéressantes de \hat{Y}_{LR} (théor) et affichera elle aussi une bonne performance.

REMERCIEMENTS

Les opinions exprimées sont celles des auteurs et n'engagent en rien le U.S. Bureau of Labor Statistics. Les auteurs tiennent également à remercier le rédacteur associé et l'arbitre pour leurs commentaires constructifs; nous croyons que leurs observations ont constitué un apport important à notre communication.

BIBLIOGRAPHIE

BAILLAR, B. (1989). Information needs, surveys, and measurement errors. Dans *Panel Surveys*, (Eds. D. Kasprzyk, G. Duncan, G. Kalton, et M.P. Singh). New York: Wiley.

DURBIN, J. (1969). Inferential aspects of randomness of sample size in survey sampling. Dans *New Developments in Survey Sampling*, (Eds. N.L. Johnson et H. Smith). New York: Wiley.

ERICSON, W.A. (1969). Subjective Bayesian models in sampling finite populations. *Journal of the Royal Statistical Society B*, 31, 195-233.

FULLER, W.A. (1981). Comment on an empirical study of the ratio estimator and estimators of its variance by R.M. Royall and W.G. Cumberland. *Journal of the American Statistical Association*, 76, 78-80.

HANSEN, M.H., HURWITZ, W.N., et MADOW, W.G. (1953). *Sample Survey Methods and Theory*, Vol. 1. New York: John Wiley and Sons.

venant de l'échantillon s . Dans le cas de la population de la CPS et de la population artificielle, les résultats donnés par l'estimateur de Horvitz-Thompson et l'estimateur par quotient étaient presque identiques, de sorte que nous présentons seulement les premiers. Sur l'ensemble des échantillons, le biais de chacun des estimateurs était négligable. Comme la théorie le laissait prévoir, $\hat{Y}_{LR}(\text{théor})$ était le plus précis des choix en présence, bien que le gain le plus élevé comparativement à \hat{Y}_{PS} n'ait été que de 4,7%, pour la population artificielle. La nécessité d'estimer H destabilise l'estimateur de régression, comme le montrent les résultats relatifs à $\hat{Y}_{LR}(\text{emp})$. Pour les populations de la NHIS et de la CPS, $\hat{Y}_{LR}(\text{emp})$ a une erreur quadratique moyenne supérieure à la fois à celle de $\hat{Y}_{LR}(\text{théor})$ et à celle de \hat{Y}_{PS} . La perte la plus prononcée concerne la population de la NHIS, où l'erreur quadratique moyenne de $\hat{Y}_{LR}(\text{emp})$ est d'environ 15% supérieure aussi bien à celle de $\hat{Y}_{LR}(\text{théor})$ qu'à celle de \hat{Y}_{PS} . Ce résultat était prévisible en raison de la taille plus faible des échantillons d'UPD de la population de la NHIS, et donc de la moins grande stabilité de l'estimation de H pour cette population.

Tableau 2

Résultats des simulations pour trois populations
5,000 échantillons ont été prélevés dans chaque population

Estimateur	Biais rel. \hat{Y} (%)	$\text{rmse}(\hat{Y})$	$100 * \left[\frac{\text{rmse}(\hat{Y})}{\text{rmse}(\hat{Y}_{PS})} - 1 \right]$	
			Population de la HIS	
\hat{Y}_{HT}	.12	.141	.05	
\hat{Y}_R	.10	.141	.02	
\hat{Y}_{PS}	.11	.141	0	
$\hat{Y}_{LR}(\text{emp})$.19	.162	14.71	
$\hat{Y}_{LR}(\text{théor})$.08	.140	-.96	
Population de la CPS				
\hat{Y}_{HT}	-.01	10.25	15.8	
\hat{Y}_{PS}	0	8.85	0	
$\hat{Y}_{LR}(\text{emp})$	-.03	9.11	3.0	
$\hat{Y}_{LR}(\text{théor})$	-.01	8.79	-.6	
Population artificielle				
\hat{Y}_{HT}	.02	2.30	-2.93	
\hat{Y}_{PS}	.12	2.37	0	
$\hat{Y}_{LR}(\text{emp})$.04	2.31	-2.41	
$\hat{Y}_{LR}(\text{théor})$.02	2.26	-4.70	

Les figures 1 à 3 présentent les résultats des simulations de l'analyse conditionnelle. Les 5,000 échantillons ont été tirés selon les facteurs de biais théoriques présentes à

la section 2.2. Le tri a été fait séparément pour chacun des estimateurs de la moyenne de la population. Dans le cas des deux estimateurs de régression, qui sont théoriquement non biaisés pour de grands échantillons, le facteur de biais pour \hat{Y}_{PS} a servi de base de tri. Les échantillons triés ont alors été répartis en 25 groupes de 200 échantillons chacun, et les biais et les erreurs quadratiques moyennes empiriques ont été calculés pour chaque groupe. On a ensuite tracé, comme le montrent les figures, le graphique des résultats des groupes par rapport aux facteurs de biais théoriques. Les ensembles de points supérieurs de chaque figure sont les erreurs quadratiques moyennes empiriques des groupes, tandis que les ensembles inférieurs sont les biais empiriques. Les deux estimateurs de régression sont, comme prévu, conditionnellement non biaisés. Les autres estimateurs, toutefois, affichent des biais conditionnels importants, lesquels, dans les ensembles d'échantillons les plus extrêmes, constituent une part importante des erreurs quadratiques moyennes. Pour la population de la CPS, l'intervalle des facteurs de biais pour \hat{Y}_{HT} est tellement plus vaste (-10 à 10) que celui des autres estimateurs que nous avons omis \hat{Y}_{HT} dans le graphique à des fins de clarté. Dans le voisinage du point d'équilibre, $\hat{N} = \hat{N}$, tous les estimateurs ont une performance à peu près identique, mais en raison d'un manque de données au stade de la préparation du plan, nous ne pouvons agir sur la mesure dans laquelle un échantillon peut être proche du point d'équilibre. Le choix le plus sûr pour le contrôle du biais conditionnel est donc $\hat{Y}_{LR}(\text{emp})$. Ce résultat est semblable à celui de Valliant (1990), qui a fait observer que dans l'échantillonnage stratifié à un seul degré, qu'il soit aléatoire ou systématique, l'estimateur de régression linéaire distinct est un bon choix pour le contrôle du biais, lorsqu'est donnée la moyenne de l'échantillon d'une variable auxiliaire.

4. BASES DÉFECTUEUSES

Les biais conditionnels examinés dans les sections précédentes étaient de nature technique et mathématique. Un problème pratique important, présent dans de nombreuses enquêtes et pouvant aussi être source de biais, est causé par la piètre couverture de la population cible. Nous nous penchons sur ce problème dans la présente section.

4.1 Le problème fondamental des bases défectueuses

Dans la plupart des applications réelles, les unités élémentaires de la population ne sont pas toutes incluses dans la base de sondage. Il n'est pas rare que dans les enquêtes auprès des ménages, certains sous-groupes démographiques, notamment les minorités, soient couverts de façon incomplète par la base de sondage. Bailar (1989), par exemple, signale qu'en 1985, l'estimation tirée de l'échantillon de la CPS du nombre total de Noirs de sexe masculin âgés de 22 à 24 ans ne correspondait qu'à 73% d'une estimation indépendante de la population totale de ce groupe. Les proportions correspondantes pour les Noirs de sexe masculin des groupes d'âge 25-29 et 60-61 étaient de 80% et de 76%.

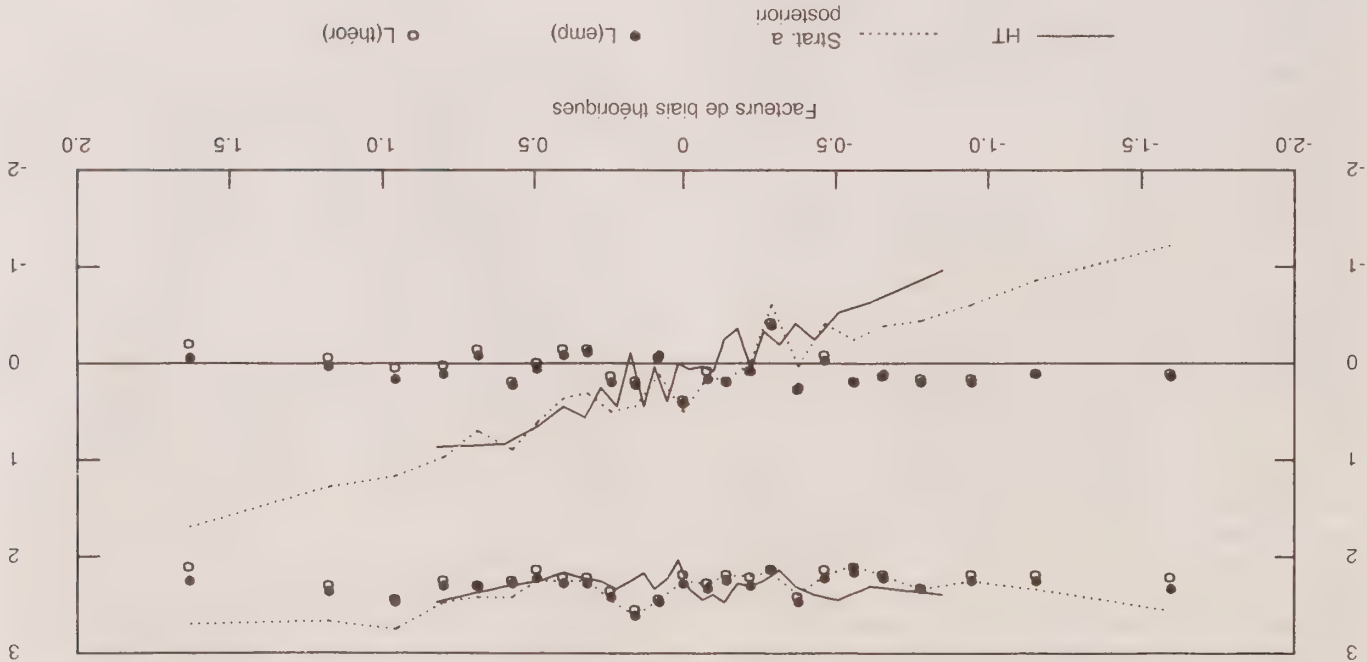


Figure 3. Simulation pour la population artificielle, $m = 200$

Tant pour la population de la CPS que pour la population artificielle, un plan d'échantillonnage stratifié à deux degrés a été utilisé. Dans le cas de la population de la CPS, des segments géographiques, employés dans l'enquête originale et composés d'environ quatre ménages voisins, ont été utilisés comme UPD, tandis que les personnes étaient les unités du deuxième degré. Dans les deux populations, 100 strates du plan ont été créées, chaque strate ayant environ le même nombre d'UPD, et un échantillon de $m = 2$ UPD a été prélevé avec probabilité proportionnelle à la taille dans chaque strate au moyen de la méthode d'échantillonnage systématique décrite par Hansen, Hurwitz et Madow (1953, p. 343). Ainsi, 200 UPD ont été sélectionnées pour les deux populations. La sélection du deuxième degré a également été semblable pour les deux populations. Dans le cas de la population de la CPS, un échantillon aléatoire simple de 4 personnes a été prélevé sans remplacement dans chaque UPD de l'échantillon pour laquelle $N_i > 4$, et toutes les personnes ont été sélectionnées dans chaque UPD de l'échantillon pour laquelle $N_i \leq 4$. Dans le cas de la population artificielle, la taille de l'échantillon à l'intérieur des UPD a été fixée à 15 au lieu de 4, ce qui a entraîné le recensement complet de la plupart des UPD de l'échantillon. Un total de 5,000 échantillons ont été prélevés dans chacune des populations pour les besoins de l'étude de simulation.

Pour chaque échantillon, nous avons calculé \hat{Y}_{HT} , \hat{Y}_R , \hat{Y}_{PS} et deux versions de \hat{Y}_{LR} . Pour la première version de

L'estimateur de régression, dénotée $\hat{Y}_{LR}(\text{emp})$ dans les tableaux, H a été estimée séparément à partir de chaque échantillon, comme cela s'imposerait en pratique. Chaque composante de Σ_{12} et de Σ_{22} a été évaluée au moyen de l'estimateur de covariance de la grappe extrême, dans une forme convenant au plan, selon la définition de Hansen et coll. (1953, p. 419). La deuxième version, dénotée $\hat{Y}_{LR}(\text{théor})$, utilisait la même valeur de H dans chaque échantillon, laquelle était une estimation d'avantage voisine de la valeur théorique de la matrice H . Dans le cas de la population de la CPS et de la population artificielle, la matrice H théorique a été estimée d'après les covariances empiriques obtenues à partir de traitements de simulation distincts de 5,000 échantillons. Dans le cas de la population de la NHIS, le plan était suffisamment simple pour permettre le calcul théorique direct de H . Quand l'échantillon d'UPD devient grand, la performance de $\hat{Y}_{LR}(\text{emp})$ devrait s'approcher de celle de $\hat{Y}_{LR}(\text{théor})$. La performance de $\hat{Y}_{LR}(\text{théor})$ est, par conséquent, représentative du mieux qu'on puisse espérer de la version empirique de l'estimateur de régression pour une taille d'échantillon donnée.

Le Tableau 2 présente les résultats de l'analyse non conditionnelle pour l'ensemble des 5,000 échantillons de chaque population. Les erreurs quadratiques moyennes ($\text{rmse}(\hat{Y}) = [\Sigma_{\hat{Y}}]^{1/2} = [\Sigma_{\hat{Y}} - \hat{Y}]^2/S]^{1/2}$, où $S = 5,000$ et \hat{Y}_s est l'une des estimations de la moyenne de la population,

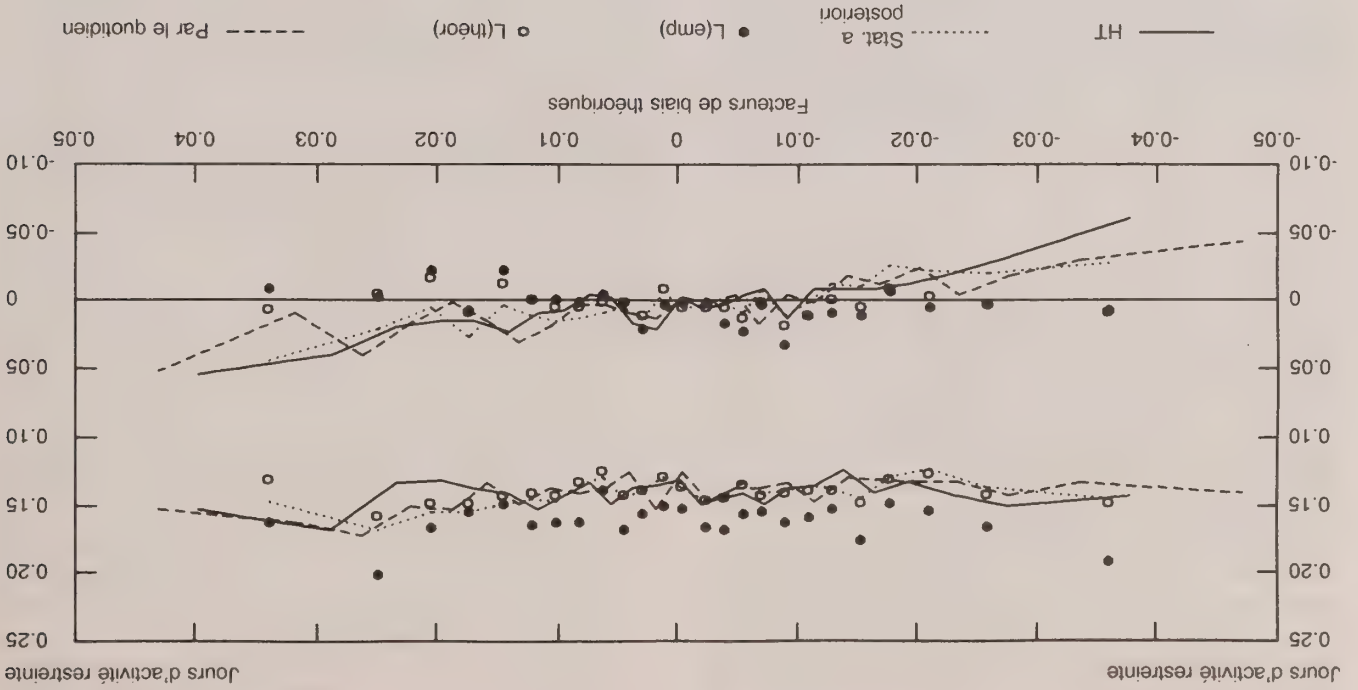


Figure 1. Simulation pour la HIS, $m = 115$

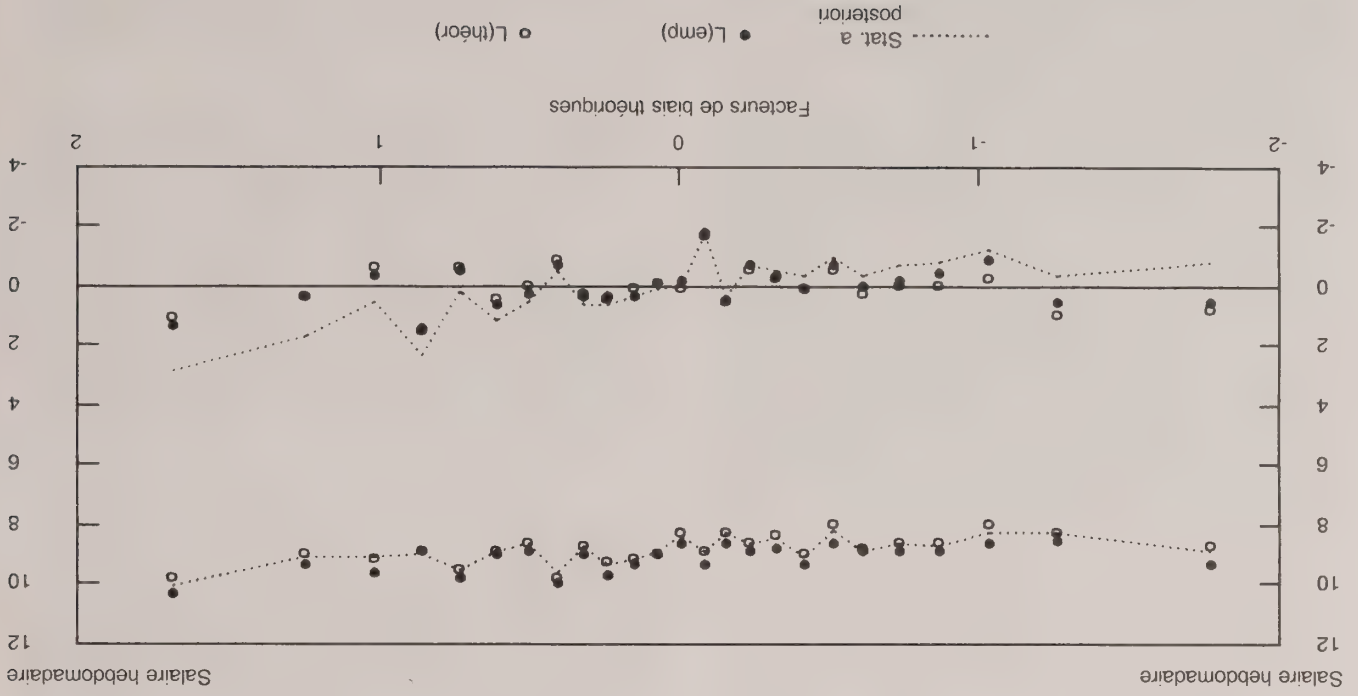


Figure 2. Simulation pour la CPS, $m = 200$

2) Estimateur par quotient:

$$E[\hat{Y}_R] = \mu + o(m^{-1})$$

$$\text{var}[\hat{Y}_R] = m^{-1}[1'[\Sigma_{11} - 2\mu\Sigma_{21} + \mu^2\Sigma_{22}]1]$$

$$+ o(m^{-(3/2)}).$$

3) Estimateur de stratification a posteriori:

$$E[\hat{Y}_{PS}] = \mu + o(m^{-1})$$

$$\text{var}[\hat{Y}_{PS}] = m^{-1}[1'[\Sigma_{11} - 2D(\mu_k)\Sigma_{21}$$

$$+ D(\mu_k)\Sigma_{22}D(\mu_k)]1] + o(m^{-(3/2)}).$$

4) Estimateur de régression linéaire:

L'espérance et la variance non conditionnelles sont les mêmes que l'espérance et la variance conditionnelles.

3. RÉSULTATS DES SIMULATIONS

La théorie élaborée dans les sections précédentes a été vérifiée dans le cadre d'un ensemble d'études de simulation fondées sur trois populations distinctes. Les tailles des populations et les paramètres de base du plan de sondage des trois études sont présentées au tableau 1. La première population est formée d'un sous-ensemble des personnes incluses dans l'échantillon du premier trimestre de la National Health Interview Survey (NHIS) de 1985, tandis que la deuxième population est constituée d'un sous-ensemble de personnes incluses dans l'échantillon de septembre 1988 de la Current Population Survey (CPS). La NHIS et la CPS sont deux enquêtes par sondage menées par le gouvernement des États-Unis. La variable d'intérêt pour la population de la NHIS est le nombre de jours d'activité restreinte au cours des deux semaines ayant précédé l'interview, tandis que la variable d'intérêt pour la population de la CPS est le salaire hebdomadaire par personne.

Des strates a posteriori, parmi les populations de la NHIS et de la CPS, ont été formées d'après des caractéristiques démographiques (comme c'est généralement le cas dans les enquêtes auprès des ménages), de façon à créer

Population	Taille de la population	Nombre d'UPD	Nombre d'UPD dans l'échantillon	Tailles des populations et paramètres de base du plan de sondage pour trois études de simulation		
				<i>N</i>	<i>M</i>	<i>m</i>
HIS	2,934	1,100	115			
CPS	10,841	2,826	200			
Artificielle	22,001	2,000	200			

Tableau 1

où N_{ik} est le nombre d'unités dans la k -ième strate a posteriori pour la i -ième UPD et α_k, β et γ sont des constantes. Plus précisément, cinq strates a posteriori ont été utilisées, avec $\alpha_k = 100k$ ($k = 1, \dots, 5$), $\beta = 10$ et $\gamma = -.05$. En tout, deux mille UPD ont été générées, et le nombre total d'unités dans la i -ième UPD, disons N_i , était une variable aléatoire de Poisson de moyenne 10. Ensuite, pour N_i , donné, les nombres d'unités dans les cinq strates a posteriori (c.-à-d. $N_{i1}, N_{i2}, \dots, N_{i5}$) pour la i -ième UPD ont été déterminés à l'aide d'une distribution multinomiale de paramètres N_i et $p_k = .20$ pour $k = 1, 2, \dots, 5$. Pour les UPD où $N_{ik} \geq 1$, la valeur de la variable d'intérêt pour la j -ième unité de la k -ième strate a posteriori dans la i -ième UPD était une réalisation de la variable aléatoire

$$Y_{ijk} = \alpha_k/N_{ik} + \beta + \gamma N_{ik} + \epsilon_{1i} + \epsilon_{2ik} + \epsilon_{3ijk}N_i.$$

$$(j = 1, \dots, N_{ik}; N_{ik} \geq 1),$$

où $\epsilon_{1i}, \epsilon_{2ik}$ et ϵ_{3ijk} sont trois variables aléatoires chi carré (à 6 degrés de liberté) réduites indépendantes. Il découle de cette structure que $E(Y_{ijk} | N_{ik})$ est donnée par (4). De plus, les valeurs de la variable d'intérêt pour les unités faisant partie d'une UPD sont corrélées, et la corrélation varie selon que les unités appartiennent ou non à la même strate a posteriori. Ce même algorithme a été utilisé pour chacune des 100 strates du plan. Vingt UPD ont été générées dans chaque strate du plan, ce qui a donné un total de 2,000 UPD.

Un plan stratifié à un seul degré a été utilisé pour la population de la NHIS, et les "ménages" ont été pris comme UPD. Dix strates du plan ont été utilisées et l'on a prélevé dans chaque strate, sans remplacement, un échantillon aléatoire simple d'environ 10% des ménages. Chaque échantillon comprenait 115 ménages et chaque ménage de l'échantillon était entièrement recensé. Un total de 5,000 échantillons de ce genre ont été prélevés pour les besoins de l'étude de simulation.

sans fournir aucun détail des calculs. Quand l'échantillon des unités du premier degré est grand, chacun des estimateurs a essentiellement la même variance conditionnelle. Les estimateurs de Horvitz-Thompson, par quotient et de stratification a posteriori sont, toutefois, conditionnellement biaisés, tandis que l'estimateur de régression linéaire ne l'est pas. Ainsi, l'estimateur de régression linéaire est celui qui affiche l'erreur quadratique moyenne asymptotique la plus faible parmi les quatre estimateurs qui nous intéressent. Rao (1992) a également signalé le caractère optimal de l'estimateur de régression à l'intérieur d'une certaine classe d'estimateurs de différences, ainsi que son biais négligeable pour de grands échantillons.

1) Estimateur de Horvitz-Thompson:

$$E[\hat{Y}_{HT} | \hat{N}] = \mu + [1'R(\hat{N} - M_2)]$$

$$\text{var}[\hat{Y}_{HT} | \hat{N}] = m^{-1} [1'(\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})1]$$

$$= m^{-1} [1'V_21] = V_{HT(c)}.$$

2) Estimateur par quotient:

$$E[\hat{Y}_R | \hat{N}] = \mu + \left(\frac{\hat{N}_{..}}{N_{..}} \right) [1'R(\hat{N} - M_2)]$$

$$= \mu + [1'R(\hat{N} - M_2)] + o(m^{-1})$$

$$\text{var}[\hat{Y}_R | \hat{N}] = (N_{..}/\hat{N}_{..})^2 V_{HT(c)}$$

$$= V_{HT(c)} + o(m^{-(3/2)}).$$

3) Estimateur de stratification a posteriori:

$$E[\hat{Y}_{PS} | \hat{N}] = \mu + [r'P(\hat{N} - M_2)]$$

$$= \mu + [1'P(\hat{N} - M_2)] + o(m^{-1})$$

$$\text{var}[\hat{Y}_{PS} | \hat{N}] = m^{-1} [r'V_r r]$$

$$= V_{HT(c)} + o(m^{-(3/2)}).$$

4) Estimateur de régression linéaire:

$$E[\hat{Y}_{LR} | \hat{N}] = \mu$$

$$\text{var}[\hat{Y}_{LR} | \hat{N}] = V_{HT(c)}.$$

Comme il a été indiqué à la section 1, de légères modifications doivent être apportées aux formules ci-dessus pour les plans, comme l'échantillonnage aléatoire simple, dans lesquels $1'N_{..} = N_{..}$. La détermination des modifications requises est directe, et nous n'en fournissons pas les détails ici.

Les biais pour de grands échantillons des trois premiers estimateurs dépendent de $N - M_2$. Autrement dit, les biais de ces estimateurs sont déterminés par l'exactitude avec laquelle l'échantillon permet d'estimer la distribution

de la population parmi les strates a posteriori. Dans certains cas spéciaux, chacun des trois premiers estimateurs peut être conditionnellement non biaisé. L'estimateur de stratification a posteriori, par exemple, sera approximativement non biaisé si $1'(H - D(\mu_k)) = 0$. Cette situation existe dans l'échantillonnage aléatoire simple et est possible, bien que certainement pas de façon générale, dans des plans plus complexes. La matrice H peut être interprétée comme la pente d'une régression multidimensionnelle de Y en N , ou de Y en N lorsque les estimations de l'échantillon sont voisines des valeurs de la population. Heureusement, en termes de superpopulation, si $E_{\xi}(Y_{ik}) = \mu_k N_{ik}$, comme dans Valliant (1993), et si E_{ξ} dénote une espérance à l'égard du modèle, alors $E_{\xi}(Y_k) = \mu_k N_k$. La pente de la régression de Y_k en N_k est alors μ_k et, dans le cas habituel où les Y_{ik} sont indépendants, H est diagonale. En fait, $H = D(\mu_k)$, de sorte que le biais conditionnel selon le plan de l'estimateur de stratification a posteriori serait nul. Si, par contre, le modèle a une ordonnée à l'origine, c.-à-d., si $E_{\xi}(Y_k) = \alpha_k + \mu_k N_k$, l'estimateur de stratification a posteriori peut avoir un important biais conditionnel selon le plan. Nous raisonnerons de cette façon dans l'étude empirique de la section 3 afin de déterminer une population pour laquelle \hat{Y}_{ps} est conditionnellement biaisé.

Un raisonnement semblable axé sur le modèle peut être appliqué aux estimateurs de Horvitz-Thompson et par quotient, afin de déterminer des populations pour lesquelles les biais conditionnels selon le plan seront vraisemblablement faibles pour de grands échantillons. Supposons, comme ci-dessus, que les Y_{ik} soient indépendants. Si chaque total de strate a posteriori n'est pas relié au nombre d'unités dans la strate a posteriori, c.-à-d. une situation particulière où $E_{\xi}(Y_k)$ ne dépend pas de N_{ik} , alors \hat{Y}_{HT} est conditionnellement non biaisé selon le plan. Si $E_{\xi}(Y_k) = \mu_k N_k$, ce qui signifie que toutes les unités élémentaires de la population ont la même moyenne peu importe la strate a posteriori, \hat{Y}_R est conditionnellement non biaisé selon le plan.

2.3 Espérances et variances non conditionnelles des estimateurs

En l'absence de condition, tous les estimateurs sont approximativement non biaisés selon le plan, comme il est indiqué ci-dessous. Les tailles relatives des variances dépendent des valeurs de Σ_{12} , Σ_{22} , μ et $D(\mu_k)$. La situation est semblable à celle d'un échantillonnage aléatoire simple visant une variable d'intérêt y et une variable auxiliaire x . Dans ce cas, la possibilité que l'estimateur par quotient, $\hat{Y}_S X/\hat{X}_S$, ou l'estimateur de régression, $\hat{Y}_S + b(X - \hat{X}_S)$, ait une moindre variance selon le plan dépend également des valeurs de certains paramètres de la population.

1) Estimateur de Horvitz-Thompson:

$$E[\hat{Y}_{HT}] = \mu$$

$$\text{var}[\hat{Y}_{HT}] = m^{-1} [1'\Sigma_{11}1].$$

Preuve: Ce résultat est analogue au résultat pour $K = 1$ présenté par Robinson (1987) et découle directement du fait que le vecteur aléatoire

$$m; \begin{bmatrix} \hat{Y} - M_1 - \Sigma_{12} \Sigma_{22}^{-1} (\hat{N} - M_2) \\ \hat{N} - M_2 \end{bmatrix}$$

tend, en distribution, vers

$$N \begin{pmatrix} \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} V_c \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ \Sigma_{22} \end{bmatrix} \end{pmatrix}.$$

En termes stricts, comme dans Robinson, nous considérons la distribution conditionnelle de \hat{Y} pour \hat{N} dans une cellule de taille $m^{-1/2}$ pour un petit ϵ . Notons que dans certains plans d'échantillonnage $1'N = N_{..}$ (par exemple ceux dans lesquels un nombre fixe d'unités élémentaires sont sélectionnées avec probabilités égales), auquel cas Σ_{22}^{-1} n'existe pas; on ne considère alors, pour l'énoncé conditionnel, que les $K - 1$ premières strates a posteriori. Dans la section qui suit, la moyenne asymptotique de \hat{Y} est utilisée comme source d'un estimateur de régression linéaire de la moyenne des y pour la population.

2. PROPRIÉTÉS CONDITIONNELLES D'ESTIMATEURS DE LA MOYENNE DE LA POPULATION

2.1 Estimateurs de la moyenne de la population

La moyenne de la population est, par définition,

$$\mu = \lim_{L \rightarrow \infty} (Y_{..}/N_{..}) = \lim_{L \rightarrow \infty} (1'Y/1'N) = \sum_{k=1}^K \phi_k \mu_k$$

où $1'$ est un vecteur ligne comprenant K fois le chiffre un. Notons que la moyenne μ n'est pas un paramètre de population finie, mais plutôt une valeur limite. Dans les grandes populations ($L \rightarrow \infty$), μ et la moyenne réelle de population seront arbitrairement voisines l'une de l'autre. Quatre estimateurs de la moyenne de la population seront examinés. Les trois premiers sont des estimateurs courants qu'on trouve dans la littérature, tandis que le quatrième est un nouvel estimateur dont la construction est suggérée par la normalité conjointe asymptotique de \hat{Y} et \hat{N} :

1) Estimateur de Horvitz-Thompson

$$\hat{Y}_{HT} = 1' \hat{Y} / 1' N = 1' \hat{Y}.$$

2) Estimateur par quotient

$$\hat{Y}_R = 1' \hat{Y} / 1' N = 1' \hat{Y} / 1' \hat{N}.$$

où $D(\mu) = \text{diag}(\mu_1, \dots, \mu_K)$ et $D(\mu_k) = \text{diag}(\mu_1, \dots, \mu_k)$ sont des matrices diagonales $K \times K$. Nous présentons ci-dessous la moyenne et la variance des quatre estimateurs

$$R = H - D(\mu), \quad \text{et} \\ P = H - D(\mu_k),$$

$$H = \Sigma_{12} \Sigma_{22}^{-1},$$

trois matrices suivantes:

À l'aide des expressions asymptotiques données plus haut, les espérances et les variances conditionnelles, pour \hat{N} donné, des quatre estimateurs peuvent être calculées. Dans le cas de la stratification a posteriori, le choix de $N_{..}$ comme condition donnée dans un plan complexe est un prolongement naturel du choix des tailles d'échantillon réalisées des strates a posteriori en vertu d'un échantillonage aléatoire simple. Dans d'autres situations, toutefois, il peut se révéler difficile de déterminer quel choisir comme condition donnée, et la réponse peut ne pas être unique (voir, p. ex., Kiefer 1977). Premièrement, définissons les

2.2 Espérances et variances conditionnelles des estimateurs

L'estimateur de régression linéaire s'appuie sur la forme de la moyenne pour grands échantillons de la variable aléatoire conditionnelle $\hat{Y} | \hat{N}$ présentée à la fin de la section 1.4, et est très semblable à l'estimateur de régression généralisée examiné par Särndal, Swensson et Wretman (1992). L'estimateur de régression linéaire (4) a également été examiné par Rao (1992) dans le contexte de l'estimation d'étalement. Il est à remarquer que l'estimateur par quotient n'exige pas que les $N_{..k}$ ou leur somme $N_{..}$ soient connus. L'estimateur de Horvitz-Thompson exige seulement que $N_{..}$ soit connu, tandis que les estimateurs de stratification a posteriori et de régression linéaire exigent que $\{N_{..k} | k = 1, \dots, K\}$ soit connu. En pratique, l'estimateur de régression linéaire présente une complication additionnelle du fait que les matrices de covariance Σ_{12} et Σ_{22} sont inconnues et doivent être estimées à partir de l'échantillon. Dans la mise en application de \hat{Y}_{LR} à la section 3, les quantités connues des populations finies ($1/N_{..}$) seront utilisées en remplacement du vecteur limite M_2 .

4) Estimateur de régression linéaire

$$\hat{Y}_{LR} = [1' (Y - \Sigma_{12} \Sigma_{22}^{-1} (\hat{N} - M_2))].$$

$$r' = [N_{..1}/N_{..1}, \dots, N_{..K}/N_{..K}].$$

où

$$\hat{Y}_{PS} = N_{..1} \sum_{k=1}^K \left(\frac{N_{..k}}{N_{..k}} \right) Y_{..k} = r' \hat{Y}$$

3) Estimateur de stratification a posteriori

qu'il y a M UPD dans la base de sondage et qu'elles sont dénotées 1, 2, ... M . Nous supposons aussi que les unités de la population peuvent être réparties en K "strates a posteriori" qui peuvent servir à des fins d'estimation. Soit y la valeur de la caractéristique d'intérêt (p. ex. revenu hebdomadaire, nombre d'heures travaillées la semaine précédente, jours d'activité restreinte au cours des deux dernières semaines, etc.) pour une unité élémentaire. À la i -ième UPD sont associés $2K$ nombres réels:

y_{ik} = somme des valeurs y pour les unités élémentaires de la i -ième UPD se trouvant dans la k -ième strate a posteriori,

N_{ik} = nombre d'unités élémentaires de la i -ième UPD se trouvant dans la k -ième strate a posteriori.

Pour chaque strate a posteriori, nous définissons ensuite:

$Y_k = \sum_{i=1}^M y_{ik}$ = somme des valeurs y pour l'ensemble des unités élémentaires dans la k -ième strate a posteriori,

$N_k = \sum_{i=1}^M N_{ik}$ = nombre total d'unités élémentaires dans la k -ième strate a posteriori.

Nous supposons, dans ce qui suit, que les N_k sont des valeurs fixes connues. Dans certaines enquêtes, les N_k peuvent en fait être eux-mêmes des estimations, mais notre analyse considère comme donné l'ensemble de N_k utilisés dans l'estimation. Dans la Current Population Survey des États-Unis, par exemple, chaque N_k est un nombre de personnes projeté d'après le recensement décennal précédent au moyen de méthodes démographiques. La somme pour la population des valeurs y est donnée par $Y_{..} = \sum_{k=1}^K Y_k$ et la taille totale de la population, par $N_{..} = \sum_{k=1}^K N_k$. Dans les sections 1 à 3, nous supposons que la base de sondage "couvre" la totalité de la population cible. À la section 4, nous examinons le problème posé par une base d'échantillonnage, c.-à-d. une base dont la couverture ne correspond pas à l'ensemble de la population cible.

1.3 Plan de sondage et estimation de base

Supposons que la base de sondage du premier degré soit divisée en L strates et qu'un plan stratifié à plusieurs degrés soit utilisé, avec un échantillon total de m UPD. Dans ce qui suit, nous avons supprimé l'indice représentant les strates du plan, de manière à simplifier la notation. Pour de définir explicitement les méthodes d'échantillonnage et d'estimation utilisées au deuxième degré et aux degrés suivants du plan. Toutefois, pour chaque UPD de l'échantillon, nous avons besoin d'estimateurs y_{ik} et N_{ik} tels que $E_2+[y_{ik}] = y_{ik}$ et $E_2+[N_{ik}] = N_{ik}$, où la notation E_2 indique l'espérance selon le plan pour le deuxième degré et les degrés supérieurs. Soit π_i la probabilité que la i -ième UPD soit incluse dans l'échantillon et $w_i = 1/\pi_i$; il s'ensuit que l'estimateur $Y_k = \sum_{i=1}^m w_i y_{ik}$ est non biaisé pour Y_k et que l'estimateur $N_k = \sum_{i=1}^m w_i N_{ik}$ est non biaisé pour N_k .

Conformément à l'approche de Krewski et Rao (1981), dans la présente section à des échantillons complexes. Nous pouvons établir nos résultats asymptotiques quand $L \rightarrow \infty$ dans un cadre formé d'une séquence de populations finies $\{\Pi_L\}$ avec L strates dans Π_L . Il est à souligner que nous supposons implicitement (sans énoncé formel) l'existence des conditions précisées dans Krewski et Rao, et développées davantage dans Rao et Wu (1985), en ce qui concerne le plan de sondage et la régularité. Les détails des preuves apportent peu de nouveaux éléments par rapport à ce qu'on trouve dans la littérature, et nous les avons omis. En notation matricielle, nous posons $Y = [Y_1 \dots Y_K]'$, $N = [N_1 \dots N_K]'$, $\bar{Y} = [Y_1 \dots Y_K]'$, $\bar{N} = [N_1 \dots N_K]'$, $\bar{Y}^* = [Y_1^* \dots Y_K^*]'$, $\bar{N}^* = [N_1^* \dots N_K^*]'$, et $V = \text{var}\{[\bar{Y}^* \bar{N}^*]'\}$, où $\bar{Y}^* = (1/N_{..}) \bar{Y}$ et $\bar{N}^* = (1/N_{..}) \bar{N}$. Notons que \bar{Y} , qui comprend $N_{..}$ au dénominateur, est utilisé pour faciliter la notation, et n'est pas une estimation des moyennes pour les strates. Comme le faisaient Krewski et Rao (1981) dans leurs conditions C4 et C5, nous supposons que:

$$\lim_{L \rightarrow \infty} \frac{Y_k}{N_k} = \mu_k, \text{ pour } k = 1, 2, \dots, K, \quad (1)$$

$$\lim_{L \rightarrow \infty} \frac{N_k}{N_{..}} = \phi_k > 0 \text{ pour } k = 1, 2, \dots, K, \text{ et } (2)$$

$$\lim_{L \rightarrow \infty} mV = V = \begin{bmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{bmatrix} \text{ (définie positive), } (3)$$

où V est partitionnée de la manière évidente. Signalons qu'encore une fois, nous avons supprimé l'indice représentant les strates du plan. Les hypothèses (1) à (3) exigent simplement que certaines quantités clés se stabilisent dans de grandes populations. La condition (2), en particulier, assure qu'aucune strate a posteriori n'est vide à mesure que s'accroît la taille de la population. Nous énonçons maintenant ce qui suit:

Résultat: Sous réserve des conditions précisées par Krewski et Rao concernant le plan de sondage et la régularité et de l'existence de V , alors, pour N donné, la distribution conditionnelle de \bar{Y} est asymptotiquement $\mathcal{N}(M_1 + \Sigma_{12} \Sigma_{22}^{-1} (N - \bar{M}_2), \Sigma_{22}^{-1} V)$, où $V = \Sigma_{11} - \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{21}$, $M_1 = \lim_{L \rightarrow \infty} \bar{Y} = [\phi_1 \mu_1 \dots \phi_K \mu_K]'$ et $M_2 = \lim_{L \rightarrow \infty} \bar{N} = [\phi_1 \dots \phi_K]'$.

1.4 Résultat asymptotique comme celui de Robinson

Propriétés conditionnelles des estimateurs de stratification a posteriori selon la théorie normale

ROBERT J. CASADY et RICHARD VALLIANT¹

RÉSUMÉ

La stratification a posteriori est une technique courante d'amélioration de la précision des estimateurs, qui consiste à utiliser des éléments d'information qui n'étaient pas disponibles au moment de la préparation du plan de l'enquête. Pour des échantillons vastes et complexes, le vecteur des estimateurs de Horvitz-Thompson des variables d'intérêt de l'enquête et des tailles de la population des strates a posteriori suivra approximativement, dans des conditions appropriées, une distribution normale multidimensionnelle. Cette normalité pour de grands échantillons amène à définir un nouvel estimateur de régression fondé sur une stratification a posteriori, analogue à l'estimateur de régression linéaire dans le cas de l'échantillonnage aléatoire simple. Nous calculons, pour de grands échantillons, le biais et les erreurs quadratiques moyennes selon le plan de ce nouvel estimateur, de l'estimateur de stratification a posteriori courant, de l'estimateur de Horvitz-Thompson et d'un estimateur par quotient. Nous utilisons des populations réelles et une population artificielle pour étudier empiriquement les propriétés conditionnelles et non conditionnelles des estimateurs en vertu d'un échantillonnage à plusieurs degrés.

MOTS CLÉS: Normalité asymptotique; estimateur de régression; bases de sondage défectueuses; estimateur par quotient; estimateur de Horvitz-Thompson.

1. INTRODUCTION

1.1 Généralités

Les études sur la théorie de l'échantillonnage au cours des vingt dernières années ont été axées dans une large mesure sur la recherche de moyens de restreindre l'ensemble d'échantillons servant de base à l'inférence. Dans une approche fondée purement sur le plan de sondage, comme le décrivent Hansen, Madow et Tepping (1983), aucune restriction de ce genre n'est imposée. Les propriétés statistiques sont calculées sous forme de moyennes sur l'ensemble des échantillons qui auraient pu être sélectionnés en vertu d'un plan particulier. Bien qu'on admette généralement qu'une certaine forme d'inférence conditionnelle, fondée sur le plan, soit souhaitable (Fuller 1981, Rao 1985, Hidiroglou et Särndal 1989), aucune théorie satisfaisante n'a encore été élaborée, sauf dans des cas relativement simples. Il existe d'autres solutions comme la théorie de la prédiction, mise au point par Royall (1971) et de nombreux autres, ainsi que l'approche bayésienne, décrite par Ericson (1969), qui évite d'établir des moyennes sur des échantillons répétés, grâce à l'utilisation de modèles de superpopulation. Une méthode d'inférence conditionnelle fondée sur le plan a été présentée par Robinson (1987) pour le cas particulier des estimations par quotient dans les enquêtes par sondage. Robinson a utilisé la théorie des grands échantillons et la normalité approximative de certaines statistiques pour produire une théorie d'inférence conditionnelle, fondée sur le plan, pour l'estimateur par quotient.

Dans la présente communication, nous étendons ce genre de raisonnement au problème de la stratification

1.2 Définitions de base et notation

La **population cible** est un ensemble bien défini d'unités élémentaires (ou analytiques). Dans de nombreuses applications, les unités élémentaires sont des personnes ou des établissements. Nous supposons que la population cible a été divisée en **unités d'échantillonnage du premier degré** (UPD). Pour les enquêtes auprès des personnes, les UPD sont en général des ménages, des groupes de ménages ou même des comtés, tandis que pour les enquêtes auprès des établissements, il n'est pas rare que l'établissement individuel soit une UPD. Quoi qu'il en soit, l'ensemble des UPD sera désigné comme la **base de sondage du premier degré** (ou simplement la **base de sondage**). Nous supposons

a posteriori. Durbin (1969), Holt et Smith (1979) et Yates (1960) ont fait valoir de façon convaincante que des échantillons de stratification a posteriori devraient faire l'objet d'une analyse conditionnelle, c.-à-d. en fonction de la distribution d'échantillonnage des unités parmi les strates a posteriori. Toutefois, comme l'a signalé Rao (1985), les difficultés de formuler une théorie exacte, fondée sur le plan, visant des échantillons finis, à l'égard de la stratification a posteriori dans les plans d'échantillonnage généraux pourraient être insurmontables. Des analyses conditionnelles, fondées sur le modèle, d'échantillons de stratification a posteriori sont présentées dans Little (1991) et Valliant (1993). La voie empruntée ici est fondée sur le plan et utilise les grands échantillons et la normalité approximative d'une manière semblable à celle de Robinson (1987) afin de permettre l'étude des propriétés conditionnelles des estimateurs.

¹ Robert J. Casady et Richard Valliant, U.S. Bureau of Labor Statistics, 2 Massachusetts Ave. N.E., Washington D.C., 20212-0001.

5.2 Problèmes liés à la codification

La seconde étape d'élaboration des données est dite opération COLIBRI (pour Codification en Ligne des Bulletins du Recensement des Individus). Recevant des bulletins toujours groupés en districts, les opérateurs et opératrices des Directions Régionales de l'INSEE procèdent à leur codification pour constituer le sondage au quart.

Physiquement, chaque opérateur (trice) travaille devant un écran qui lui indique l'identifiant du prochain logement à inclure dans l'échantillon au quart dont il doit codifier tous les BI.

Le contrôle de la qualité de la codification est également réalisé par sondage. L'unité de contrôle est l'ensemble du travail réalisé en une semaine dans une Direction Régionale. L'opération dure un peu plus d'un an dans les 22 Directions Régionales soit plus de mille sondages. L'unité à contrôler est le ménage (c'est-à-dire l'ensemble des BI d'un ménage tiré pour figurer dans l'échantillon de contrôle). L'objectif est d'estimer la proportion de bulletins comportant une erreur. Pour cela, on détecte automatiquement ceux pour lesquels apparaît une divergence entre les deux codifications. Une opération de réconciliation permet de chiffrer le nombre d'erreurs. La théorie de ce contrôle a fait l'objet de la partie 4 de cet article. L'indice de difficulté des bulletins a été élaboré à partir des données déjà saisies pour une étude faite à partir du précédent recensement et d'un test. Les modalités pratiques et les enseignements tirés de ces contrôles sont détaillés dans G. Badeyan (1992).

L'application pratique et numérique de la théorie repose sur des hypothèses concernant les ordres de grandeurs des différents paramètres (ce qui demande qu'on puisse les raccrocher à une interprétation physique simple). Dans la phase de préparation du recensement, sans mesures préalables très précises, on a utilisé les valeurs $\sigma/a = 0,5$ et $C_1/C_0 = 0,1$.

À la suite de diverses hypothèses sur les autres paramètres et de discussions entre experts, il a été décidé un contrôle portant sur 50 districts chacun d'eux étant contrôlé pour environ 20 BI (par région et par semaine). Cet ordre de grandeur initial pouvait, évidemment, être modulé dans la suite du contrôle, les paramètres du modèle pouvant être réestimés après chacun d'eux.

Remarque finale:

Ce problème fait apparaître des résultats un peu surprenants sur lesquels il est utile de réfléchir un peu.

Dans un premier cas, nous avons supposé qu'on pouvait isoler chaque bulletin. On tirait alors ceux-ci avec des probabilités proportionnelles à leur difficulté individuelle. On supposait, dans une certaine mesure, que le coût d'utilisation de l'information individuelle était nul.

REMERCIEMENTS

L'auteur tient à remercier de leurs commentaires très positifs le rédacteur en chef, le rédacteur associé et l'arbitre qui ont examiné ce travail. Il en va de même de Claude Thelot dont les remarques sont à l'origine d'un certain nombre de développements et de Gérard Badeyan qui a mis en place à l'INSEE les techniques ici préconisées. Il remercie Française Hittier sans qui ce texte n'aurait jamais pu exister.

BIBLIOGRAPHIE

- BADÉYAN, G. (1992). Communication aux secondes Journées de Méthodologie Statistique, 17 et 18 juin 1992, INSEE, Paris.
- COCHRAN, W. (1977). *Sampling Techniques*, (3^{ème} édition). New York: Wiley.
- DESABIE, J. (1965). *Théorie et Pratique des Sondages*. Paris: Dunod.
- LUENBERGER, D.G. (1973) *Introduction To linear and Non-linear Programming*. New York: Addison-Wesley.
- SÄRNDAAL, C.-E., SWENSSON, B., et WRETMAN, J. (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.

Nous avons obtenu une solution complète du problème.

Remarque 1:

Dans les deux cas qui ont été traités S_k^* est multiplié par $C_{1/2}$ si les X_i sont multipliés par C . La formule qui donne n_k est donc bien invariante à l'échelle de mesure.

Remarque 2:

La solution dans le cas du tirage en grappes privilégie le tirage de petites grappes dont les unités finales on un fort indice de difficulté.

Remarque 3:

Ici comme dans les parties précédentes nous déterminons les probabilités d'inclusion simple mais pas les probabilités d'inclusion double. L'algorithme de tirage, qui fixe ces dernières, est donc sans influence. Ceci est relativement naturel si nous nous disons que l'information auxiliaire utilisée pour optimiser le tirage déterminera les π_k et $\pi_{1/k}$ mais ne peut pas avoir d'influence sur les probabilités doubles.

5. APPLICATIONS AUX CONTRÔLES PAR SONDAGE DE LA QUALITÉ DU RECENSEMENT DE 1990 EN FRANCE

5.1 Problème de contrôle de la saisie

Les techniques d'échantillonnage décrites aux paragraphes 2 et 3 ont été motivées par la nécessité de contrôler la saisie du recensement de 1990. Pour comprendre la nature des problèmes statistiques une description des principes d'exploitation est utile.

L'unité de base pour la collecte est le district, correspondant, en ville, à un pâté de maisons et, à la campagne, à un village ou une réunion de hameaux. Il peut comporter une population variant de zéro à environ deux mille habitants. La moyenne est de 150 logements pour 350 habitants environ.

À mesure de l'achèvement de la collecte et de sa vérification, les différents bulletins du recensement, notamment les bulletins individuels (BI) et les feuilles de logement (FL), sont soigneusement comptés pour chaque district. Les données récapitulatives des districts sont saisies sur support informatique, tandis que les bulletins groupés, dans une chemise de district, partent pour la saisie. Des ensembles de districts groupant environ 100,000 logements sont constitués. Ce sont les unités de traitement (UT). Chaque UT est saisie par une entreprise à façon pour le compte de l'INSEE.

L'INSEE, le "client" en termes de théorie du contrôle, vérifie la qualité du travail de chaque façonier en consultant par sondage un certain nombre de bulletins dans chaque UT.

Le but du sondage décrit au paragraphe 2 est d'estimer la proportion de bulletins erronés dans chaque UT avec une précision (écart-typé) de un point. La proportion maximum de bulletins erronés ne saurait excéder 4%. Un test de recensement portant sur environ 400 districts permet d'estimer les valeurs des deux paramètres du modèle. On trouve:

$$\sigma^2 \approx P^2 \approx 14.10^{-4}$$
$$\tau^2 \approx P \approx 4.10^{-2}$$

La fonction de coût (1.1) a pu être évaluée en temps de

travail. Les mesures faites dans les ateliers ont permis d'estimer à 5 minutes le temps de manipulation d'une chemise de district (du moment où on va la chercher dans une étagère au moment où on la range) et à 30 secondes le temps de saisie d'un BI. Avec des données numériques, l'optimisation du plan avec les hypothèses du paragraphes 1 conduit à contrôler 40 districts par lot de traitement et 16 bulletins à l'intérieur de chaque district.

Après discussion de cette solution avec l'équipe responsable du recensement, il est apparu qu'il fallait, en fait, contrôler deux types de documents: les bulletins individuels (BI) et les feuilles de logement (FL). On avait été conduit à négliger ces dernières, en première approximation, parce qu'elles sont moins susceptibles de receler des erreurs et que leur temps de codification est plus court (la moitié environ) que celui nécessaire pour un BI. Toutefois, dans certains districts, par exemple dans les communes très touristiques, on trouve une forte majorité de résidences secondaires et donc beaucoup de FL pour très peu de BI. Cette situation demande une étude particulière, dont la théorie a été faite au paragraphe 3.

Dans le cas du recensement le nombre G de groupes vaut 2 ($g = 1$ pour les BI et $g = 2$ pour les FL). Les données numériques relatives aux deux groupes étaient les suivantes:

$$\begin{aligned} & \cdot P_1 = 0,04 \quad \sigma_1 = P_1 \quad \tau_1^2 = P_1(1 - P_1) - \sigma_1^2 = P_1 - 2P_1^2 \\ & \cdot P_2 = 0,01 \quad \sigma_2 = P_2 \quad \tau_2^2 = P_2 - P_2^2 \\ & \cdot \psi_1 = (0,0075)^2 \quad \psi_2 = (0,0150)^2 \end{aligned}$$

Pour la fonction de coût on a pris $C_0 = 5$ minutes, $C_1 = 0,5$ minute et $C_2 = 0,25$ minute. L'optimisation du problème selon les hypothèses de la partie 3.2.b a conduit à examiner 73 districts par unité de traitement. La solution pratique à considérer consiste à traiter 15 bulletins individuels par district ainsi que les FL associées. Pour les districts comportant moins de 15 BI, l'intégralité des BI était traitée. Pour les districts vides de BI on traitait 4 FL (si ce nombre était inférieur au nombre de FL du district).

Remarque:

La technique de la partie 2 semble avoir un domaine d'application assez fréquent. Elle a, en particulier, été utilisée pour l'échantillonnage de l'enquête française de 1992 sur les migrations en ce qui concerne la population de nationalité étrangère. Pour les agglomérations de moins de 20,000 habitants, l'échantillon était à deux degrés, le premier degré de sondage étant constitué par les 90 départements où ce type d'agglomération existe. La population étrangère (sur la base du recensement) était divisée en 8 groupes de nationalités pour lesquels on devrait obtenir des indicateurs ayant la même précision.

Ici, $n_k = \sum_{i \in k} \pi_i / k$ est la taille de l'échantillon tiré dans le district k (supposé de taille fixe à s_1 fixé). Son espérance vaut:

$$C_T = \sum_{k \in U_0} \pi_k (C_0 + C_1 n_k).$$

Posons

$$\pi_{i/k} = n_k P_i \text{ (avec } \sum_{i \in k} P_i = 1) \text{ et } Q_k = \pi_k n_k.$$

Le problème d'optimisation est maintenant:

$$\text{Min: } C_0 \sum \pi_k + C_1 \sum Q_k$$

$$\text{sous: } \sigma^2 \sum \frac{\pi_k}{X_k^2} + a \sum \frac{1}{X_i} \sum_{i \in k} \frac{Q_k}{P_i} \leq v_0.$$

Sous cette forme, on constate avec plaisir qu'on peut minimiser les termes en $\sum_i X_i / P_i$ indépendamment du reste. Autrement dit, n_k n'a pas d'incidence sur ce terme. Laissons l'optimisation du second degré de tirage pour plus tard et notons seulement S_k^* la valeur optimisée de $\sum_i X_i / P_i$. Avec un multiplicateur de Lagrange λ on obtient par dérivation par rapport aux π_k puis au Q_k :

$$*C_0 = \lambda \sigma^2 \frac{\pi_k^2}{X_k^2} \text{ soit } \pi_k \text{ proportionnel à } X_k \quad (4.3.1)$$

$$*C_1 = \lambda a \frac{Q_k^2}{S_k^2} \text{ d'où } \pi_k = \left(\frac{C_1}{C_0} \right)^{1/2} a^{1/2} \frac{\sigma}{S_k^*} X_k \quad (4.3.2).$$

En particulier on tirera les Unités Primaires avec des probabilités proportionnelles à leur difficulté totale, ce qui est un résultat standard (voir par exemple Särndal, Swensson, Wretman 1992, chapitre 12).
Passons maintenant au tirage infra-district (deuxième degré de sondage).
Commentons par un cas simple et naïf: on tire les bulletins individuellement. La minimisation conduit à P_i proportionnelle à $\sqrt{X_i}$. Un calcul simple nous montre qu'alors $S_k^* = \sum_{i \in k} \sqrt{X_i}$. Ceci nous permet de calculer n_k grâce à (4.3.2) et notre problème est entièrement résolu.
En fait les choses sont plus compliquées. Pour des raisons assez naturelles, on ne sélectionnera que les bulletins que de ménages entiers. Autrement dit le sondage au second degré est un sondage en *grappes*. Les valeurs de P_i seront les mêmes, soit P_m , pour tous les membres d'une même grappe (ménage) m .

Notons par X_m^i la somme des X_i des individus i du ménage m . Le problème est donc de minimiser $\sum X_m^i / P_m$ sous $\sum n_m P_m = 1$ avec n_m taille du ménage m . On trouve facilement la solution:

$$P_m = \sqrt{X_m^i} / \sum n_m \sqrt{X_m^i},$$

avec $X_m^i = X_m / n_m$, difficulté moyenne des bulletins du ménage m . Par suite on trouve $S_k^* = \sum n_m \sqrt{X_m^i}$.

Cette solution nous permet de déterminer le nombre n_k d'unités finales à tirer grâce à (4.3.2). Le nombre de grappes (*ménages*), en revanche n'est pas déterminé. Cette difficulté était prévisible. La fonction de coût, en effet, ne fait pas intervenir cette contrainte. Pour obtenir le nombre n_k de grappes à tirer, on s'arrangera de façon à ce que l'espérance du nombre d'unités finales soit égale à n_k . Elle vaut:

$$m_k \left(\sum n_m \sqrt{X_m^i} \right) / \sum \sqrt{X_m^i}$$

d'où:

$$m_k = n_k \frac{\sum \sqrt{X_m^i}}{\sum n_m \sqrt{X_m^i}}.$$

Compte tenu de (4.3.2) on a aussi:

$$m_k = \left(\frac{C_1}{C_0} \right)^{1/2} a^{1/2} \frac{\sigma}{S_k^*} \frac{X_k}{\sum \sqrt{X_m^i}}.$$

et la probabilité de tirer un ménage vaut alors:

$$\frac{\sum \sqrt{X_m^i}}{m_k \sqrt{X_m^i}}.$$

Après quelques manipulations algébriques, on trouve la valeur de la variance optimale:

$$E_k \text{ Var} (aX)^{\text{OPT}} = \frac{m}{(\sigma X)^2} \left(1 + \frac{\sigma}{a} a^{-1/2} S^* \left(\frac{C_0}{C_1} \right)^{1/2} \right).$$

Cette forme respecte le caractère homogène des différents facteurs. On a, en particulier $a^{-1/2} S^* / X = a^{1/2} S^* / aX$, le dénominateur est interprétable comme un nombre total d'erreurs dans un lot, tandis que le numérateur est homogène à une taille.

Pour le second terme, on a, *conditionnellement* à s_1 :

$$\text{Var} \left(\frac{\sum_i a_k X_i}{\sum_i X_i} \middle| s_1 \right) = \left(\sum_i \frac{\pi_k}{X_i} \right)^{-2} \cdot \sum_i \text{Var}(\hat{a}_k) \frac{\pi_k^2}{X_i^2}.$$

L'espérance de cette quantité vaut approximativement

$$X^{-2} \sum_i E \text{Var}(\hat{a}_k | s_1) \frac{\pi_k^2}{X_i^2}, \quad (4.1.2)$$

avec:

$$\text{Var}(\hat{a}_k | s_1) = \text{Var} \frac{\sum_i \frac{X_i}{Y_i - a_k X_i} \sum_{s_k} \frac{\pi_{i|k}}{X_i}}{\sum_{s_k} \frac{\pi_{i|k}}{X_i}} \approx \frac{1}{\text{Var} \sum_{s_k} \frac{\pi_{i|k}}{Y_i - a_k X_i}} \sum_{s_k} \frac{\pi_{i|k}}{X_i^2}.$$

$$= \frac{1}{\sum_{i \in k} \left(\frac{X_i^2}{(Y_i - a_k X_i)^2} \pi_{i|k} \right)} + \sum_{k \neq i} \frac{\sum_{i' \neq k} \pi_{i'k} \pi_{i|k}}{(Y_i - a_k X_i)(Y_j - a_k X_j) \pi_{ij|k}}.$$

Comme dans les parties précédentes, nous arrivons à des formules complexes et, finalement, inutilisables. Un modèle va nous simplifier un peu l'existence.

4.2 Intervention d'un modèle

Il aura la même structure que ceux qui ont déjà servi antérieurement:

a) Les a_k seront des variables aléatoires indépendantes de même espérance et de même variance:

$$E_{\xi} a_k = a \quad \text{Var}_{\xi} a_k = \sigma^2.$$

La variance prend en compte l'influence de l'opérateur ou de l'opérateur, qu'on renonce à isoler, mais aussi celle du jour de la semaine, de l'heure dans la journée, de certains jours du mois etc. . . .

b) Conditionnellement à a_k , les Y_i de l'Unité Primaire k sont des variables de Bernoulli indépendantes avec $E_{\xi}(Y_i | k) = a_k X_i$

$$\text{Var}_{\xi}(Y_i | k) = a_k X_i - a_k^2 X_i^2.$$

Remarque:

La variable X_i n'a pas de véritable sens concret et n'est d'ailleurs définie qu'à un facteur d'échelle près. En revanche $a_k X_i$ et σX_i ont une interprétation physique invariante, car ce sont des probabilités. Dans tout ce qui suit il faudra toujours garder en tête que les résultats devront être invariants si les X_i sont multipliés par un facteur arbitraire à condition que a et σ soient divisés par le même facteur. En particulier $\text{Var}(\hat{a})$ n'a pas de sens "concret".

Comme précédemment nous allons étudier la variance anticipée, espérance sous modèle de la somme de (4.1.1) et (4.1.2).

Pour le premier terme, l'espérance des produits croisés est nulle, comme de bien entendu. L'espérance sous modèle de ce terme est donc:

$$X^{-2} \sigma^2 \sum_k \frac{\pi_k}{X_k^2}.$$

Pour le second terme on trouve: (vu les définitions données au 4.2.a et 4.2.b):

$$X^{-2} \sum_k \frac{\pi_k}{X_k^2} \cdot \frac{1}{\sum_i \frac{X_i^2}{(a_k X_i - a_k^2 X_i^2)} E_{\xi} \frac{\pi_{i|k}}{X_i^2}} = X^{-2} \sum_k \frac{\pi_k}{1 \sum_i a X_i - (a^2 + \sigma^2) X_i^2} \pi_{i|k}.$$

Globalement donc:

$$E_{\xi} \text{Var}(\hat{a} X) = \sigma^2 \sum_{k \in U_0} \frac{\pi_k}{X_k^2}$$

Ici pas de miracle algébrique. Pour simplifier nous admettons que $(a^2 + \sigma^2) X_i^2$ est négligeable devant $a X_i$. Numériquement on peut attendre $a X_i = 2 \text{ à } 5 \times 10^{-2}$ et $(a^2 + \sigma^2) X_i^2 = 3 \text{ à } 30 \times 10^{-4}$. D'où notre approximation

$$E_{\xi} \text{Var}(\hat{a} X) \approx \sigma^2 \sum_{k \in U_0} \frac{\pi_k}{X_k^2} + a \sum_{k \in U_0} \frac{1}{\sum_{i \in k} \pi_{i|k}} \frac{\pi_{i|k}}{X_i}.$$

4.3 Optimisation du plan de sondage

Nous utiliserons la fonction de coût suivante:

$$C = \sum_{s_1} (C_0 + C_1 m_k).$$

Etape 3: Les π_k sont déterminées par les relations (3.1.8). La somme des π_k fixe, en particulier, le nombre d'UP à tirer.

Etape 4: Les n_{kg} sont déterminées par les relations (3.1.4). On peut ensuite itérer par retour à l'étape 2 en espérant que cet algorithme converge vers la solution d'optimisation.

Remarque: La probabilité de tirer une unité de type g vaut

$$\pi_k n_{kg} / N_{+g} = \left(\frac{C_g}{\lambda_g} \right)^{1/2} \tau_g / N_{+g}.$$

Elle ne dépend pas de l'Unité Primaire k et est donc la même pour chaque unité d'un groupe g donné (sondage à probabilités égales). On en déduit la taille n_{+g} de l'échantillon dans le groupe g , ou du moins, son espérance mathématique. Pratiquement, il arrive que l'on fixe "autocritériquement" les tailles des échantillons. Ceci revient à déterminer les λ_g , ou, implicitement, des variances τ_g . Ce résultat est assez naturel lui aussi.

4. ESTIMATION OPTIMALE À L'AIDE D'UNE MESURE DE LA DIFFICULTÉ À CODER UN ENREGISTREMENT

Il s'agit d'estimer la proportion de bulletins présentant une erreur de codification dans l'"univers" U de tous les bulletins codifiés une semaine donnée dans une Direction Régionale. Le caractère particulier du problème est que suivant: tous les bulletins i sont déjà précodifiés ce qui permet, grâce à des informations tirées de l'essai de recensement, d'attribuer à chacun d'eux une variable numérique positive X_i qui traduit sa "difficulté". Cette variable a été calibrée de façon à ce que Y_i (qui vaut 1 en cas d'erreur et 0 sinon) ait une "espérance" proportionnelle à X_i .

Toujours pour les mêmes raisons de coût du contrôle, on est amené à envisager un sondage à deux degrés: - au premier degré de sondage on tire un échantillon s_1 de districts k (les Unités Primaires) à probabilités égales π_k à déterminer. On notera $\pi_{k\ell}$ les probabilités d'inclusion double pour cet échantillonnage.

- au second degré de sondage, on tirera un échantillon s_k d'unités finales (les bulletins) dans l'unité primaire k . On notera $\pi_{i|k}$ la probabilité d'inclusion de l'unité dans l'unité primaire k , $\pi_{ij|k}$ la probabilité d'inclusion du couple (i, j) dans les unités primaires et $s = U_{k \in s_1} s_k$ l'échantillon d'unités finales.

On notera $X_k = \sum_{i \in k} X_i$ le total des X_i dans l'unité primaire k , $X = \sum_{k \in U_0} X_k = \sum_U X_i$ et on adoptera des notations analogues pour toutes les variables. (U_0 désigne la population des Unités Primaires - districts, U la population des Unités finales - bulletins).

4.1 Choix d'estimateur et variance

a) Au niveau de l'Unité Primaire numéro k il est naturel d'estimer le total Y_k des Y_i pour $i \in k$ par le ratio:

$$Y_k = X_k \left(\sum_{i \in k} Y_i / \pi_{i|k} \right) / \left(\sum_{i \in k} X_i / \pi_{i|k} \right) = X_k \hat{a}_k.$$

Ici \hat{a}_k estime $a_k = Y_k / X_k$ avec un faible biais.

b) Pour estimer le ratio Y/X on utilisera:

$$\hat{a} = \frac{\sum_{k \in s_1} \frac{Y_k}{\pi_k}}{\sum_{k \in s_1} \frac{X_k}{\pi_k}} = \frac{\sum_{k \in s_1} \frac{X_k}{\pi_k} \hat{a}_k}{\sum_{k \in s_1} \frac{X_k}{\pi_k}}.$$

c) Si on veut estimer R , on remarquera que:

$$R = \frac{Y}{X} \cdot \frac{X}{W},$$

où X et W sont des totaux connus (difficulté totale et nombre total de bulletins, par exemple). Comme la variable X_i a été choisie pour sa bonne corrélation avec Y_i , un estimateur *a priori* intéressant de R sera:

$$\hat{R} = \hat{a} \frac{W}{X}$$

de sorte que la seule véritable question porte sur l'estimation de $a = \sum_k a_k X_k / X$.

d) On aura:

$$\text{Var}(\hat{a}) = \text{Var}(E(\hat{a} | s_1) + E \text{Var}(\hat{a} | s_1).$$

Pour le premier terme, compte tenu du fait que \hat{a}_k estime (à peu près) sans biais a_k , on peut écrire:

$$\text{Var}(E(\hat{a} | s_1)) = \frac{1}{X^2} \text{Var} \left(\sum_{k \in s_1} \frac{X_k}{(a_k - a) X_k} \right) = \frac{1}{X^2} \sum_{k \in s_1} \frac{X_k^2}{(a_k - a)^2 X_k^2} = \sum_{k \neq l} \sum_{k \neq l} (a_k - a)(a_l - a) \frac{\pi_k \pi_l}{X_k X_l \pi_{kl}}. \quad (4.1.1)$$

Les π_k étant, pour l'instant, destinées à être connues à un facteur près, on peut écrire:

$$(3.1.4) \quad \pi_k n_{kg} = \left(\frac{C_g}{\lambda_g} \right)^{1/2} \frac{N_{kg}}{T_g}.$$

Par sommation sur k on en déduit que:

$$(3.1.5) \quad E n_{+g} = \sum^U \pi_k n_{kg} = \left(\frac{C_g}{\lambda_g} \right)^{1/2} T_g.$$

La taille totale de l'échantillon dans chaque groupe est donc directement liée au multiplicateur λ_g .

La dérivation du Lagrangien par rapport aux π_k nous donne de nouvelles relations qui se simplifient miraculeusement si on utilise aussi (3.1.4). On obtient:

$$(3.1.6) \quad C_o = \sum^g C_g \left(\frac{T_g}{\sigma_g} \right)^2 n_{kg}^2,$$

où encore, si on introduit les nombres

$$n_g^* = \left(\frac{C_o}{T_g} \right)^{1/2} \frac{\sigma_g}{T_g},$$

on écrit:

$$(3.1.7) \quad \sum^g \left(\frac{n_g^*}{n_{kg}^*} \right)^2 = 1.$$

Comme on s'en rend compte en jetant un oeil à la formule (2.2.4), les n_g^* sont les nombres d'unités secondaires à tirer par UP s'il n'y a qu'un seul groupe; n_{kg}^* sera toujours inférieur à n_g^* .

De (2.1.4), (3.1.5) et (3.1.7) on tire les relations:

$$(3.1.8) \quad \pi_g^2 = \frac{1}{C_o} \sum^g \lambda_g \sigma_g^2 \left(\frac{N_{kg}}{N_{+g}} \right)^2.$$

Ainsi, les π_k sont proportionnelles aux quantités T_k telles que $T_g^2 = \sum^g \lambda_g \sigma_g^2 N_{kg}^2 / N_{+g}^2$ qui apparaissent comme la mesure de taille adéquate. Les relations (3.1.4) montrent que, à k fixe, les n_{kg} sont proportionnelles à $n_g^* \lambda_g^{1/2} \sigma_g N_{kg} / N_{+g}$, ce qui compte tenu de (3.1.7) conduit à:

$$(3.1.9) \quad n_{kg} = n_g^* \lambda_g^{1/2} \sigma_g \frac{N_{kg}}{N_{+g}} T_k^{-1}.$$

3.2 Solutions explicites dans deux cas particuliers

a) Si les λ_g étaient connus, c'est-à-dire si on minimisait $\sum^g \lambda_g V_g$ sous une contrainte de coût, alors (3.1.2) et (3.1.9) nous permettraient de calculer les T_k . En reportant:

$$\pi_k = m T_k / T \quad T = \sum^U T_k, \quad m \text{ nombre d'unités primaires à tirer}$$

dans la contrainte de budget $C_T \leq C_T^*$, on trouve:

$$C_T^* = \frac{m}{T} \sum^U \left(C_o T_k + \sum^g C_g n_g^* \lambda_g^{1/2} \sigma_g \frac{N_{kg}}{N_{+g}} \right) \quad \text{soit:}$$

$$m = C_T^* \left(C_o + \sum^g C_g n_g^* \cdot \frac{T}{\lambda_g^{1/2} \sigma_g} \right).$$

Si un seul des λ_g est différent de zéro, on peut vérifier assez facilement qu'on retrouve le résultat donné à la fin de la section (2.2).

b) Le problème initial (min C_T sous $V_g \leq V_g^*$) se résoud assez facilement dans deux cas particuliers:

b1 - *Dispersion maximale* des groupes. Pour toute UP k , on a $N_{kg} = N_{k+}$ pour un certain k . Le problème est décomposé en G problèmes distincts, chacun d'eux étant du type étudié dans la section 2.

b2 - *Dispersion minimale*: La répartition est la même dans toutes les UP; autrement dit on a pour tout k et g

$$N_{kg} = N_{k+} \frac{N}{N_{+g}} \quad \text{avec} \quad \left(N = \sum^g N_{+g} \right),$$

T_k est alors proportionnelle à N_{k+} , et les n_{kg} sont des quantités $n_g^* u_g$ indépendantes de k .

Avec $\pi_k = m N_{k+} / N$, on obtient en écrivant $V_g = V_g^*$:

$$m V_g^* = \sigma_g^2 + T_g^2 / n_g^* u_g$$

soit:

$$m = \frac{\sigma_g^2}{T_g^2} + u_g^{-1} \frac{n_g^* V_g^*}{T_g^2}.$$

On obtient ainsi $G-1$ relations linéaires entre les u_g^{-1} ce qui permet, en principe, de résoudre complètement le problème sachant que la somme des u_g^2 vaut 1.

3.3 Un algorithme numérique permettant de trouver la solution optimale dans le cas général

Une résolution numérique itérative du problème peut se faire de la façon suivante:

Étape 1: On fixe une allocation approximative de l'échantillon dans chaque groupe, soit n_{+g} unités dans le groupe g . Pour y arriver on peut, par exemple, se servir de la solution approximative avec les hypothèses du point a) ou du point b).

Étape 2: La valeur des λ_g est déterminée par les relations (3.1.5):

$$\lambda_g = C_g n_{+g}^2 / T_g^2.$$

3. ESTIMATION OPTIMALE DANS LE CAS D'UN SONDAGE À DEUX DEGRÉS OU LES UNITÉS PRIMAIRES SONT STRATIFIÉES

La dure réalité des choses nous amène à compléter un peu le problème car on doit, en fait, contrôler indépendamment plusieurs types de bulletins. Ceci amène à poser un problème assez général qui est le suivant:

Pour chaque unité primaire (ici les districts d'une unité de traitement) on connaît les effectifs N_{kg} , d'unités secondaires appartenant à G groupes. La "population" de l'UP numéro k vaut $N_{k+} = \sum_g N_{kg}$; celle du groupe g vaut $N_{+g} = \sum_k N_{kg}$. Comme dans ce qui précède on cherche avec quelle probabilité d'inclusion π_k échantillonner l'UP numéro k , le nombre d'UP à tirer et l'allocation n_{kg} de l'échantillon parmi les différents groupes dans l'UP k , sachant que ces n_{kg} unités sont tirées par un sondage aléatoire simple parmi les N_{kg} unités tirables.

3.1 Recherche d'un plan optimal à l'aide d'un modèle

On postule, dans chacun des groupes, un modèle identique à celui formulé à la section (2.1) (ou sous une forme plus générale dans la remarque qui la termine).

Pour $g = 1$ à G on aura donc:

$$v_g = E_i \text{Var}(P_g) = N_{-2}^{-g} \sum_k \frac{\pi_k}{N_{kg}} (o_2^g + \tau_2^g/n_{kg}). \quad (3.1.1)$$

La fonction de coût est donnée par la forme générale (1.2). On va chercher à minimiser l'espérance du coût de sondage:

$$C_T = \sum_U \pi_k \left(C_o + \sum_g n_{kg} C_g \right), \quad (3.1.2)$$

sous les contraintes $V_g \leq v_g$, où les quantités v_g , sont fixées de façon extérieure, par exemple par la qualité des données qu'on veut obtenir et la rigueur du contrôle.

Sous cette forme, le problème peut s'avérer assez complexe. Nous allons écrire un "Lagrangien" général:

$$L = \lambda C_T + \sum_g \lambda_g V_g.$$

Le problème posé fixe $\lambda = 1$ et les λ_g sont des multiplicateurs à déterminer. Une variante simple consiste à fixer les λ_g : on désire alors minimiser une combinaison linéaire donnée des variances sous une contrainte de coût. Dans toutes les hypothèses, on obtient par dérivation par rapport aux n_{kg} (considérées comme des variables réelles):

$$\lambda \pi_k^2 C_g = \lambda_g N_{-2}^{-g} N_{kg}^2 \tau_2^g / n_{kg}^2. \quad (3.1.3)$$

C'est la justification habituelle d'un sondage autopondéré avec un premier degré tiré avec des probabilités proportionnelles à une mesure de taille (Voir par exemple Cochran 1977).

Comme n_k ne dépend pas de N_k on ne pourra avoir $n_k = N_k$ et $\mu_k < 0$ que si $N_k \leq n^*$. L'équation (2.2) nous permet alors d'obtenir les probabilités d'inclusion à un facteur près:

$$\pi_k = \lambda_{1/2} N_k \left(\frac{o_2^2 + \tau_2^2/N_k}{C_o + C_1 N_k} \right)^{1/2} = \lambda_{1/2} N_k^{1/2} \left(\frac{N_k C_1 + C_o}{N_k o_2^2 + \tau_2^2} \right)^{1/2}. \quad (2.2.6)$$

Les relations (2.2.5), valide si $N_k \geq n^*$ et (2.2.6) valide si $N_k \leq n^*$ établissent que π_k est proportionnelle à une variable connue $T_k = f(N_k)$ dont le graphique est donné à la figure 1.

Pour spécifier entièrement le sondage il reste à trouver le nombre m d'unités primaires à tirer. Or, $T = \sum_U T_k$ est aussi une quantité connue. En se restreignant à un échantillonnage de taille fixe on aura donc $\pi_k = m T_k / T$. On trouve m en portant cette valeur dans la contrainte de variance soit:

$$m V_o = o_2^2 + \tau_2^2 / n^*.$$

Si, en première approximation, on prend $T_k = N_k$, on obtient la formule simplifiée:

On a ainsi obtenu une solution complète au problème.

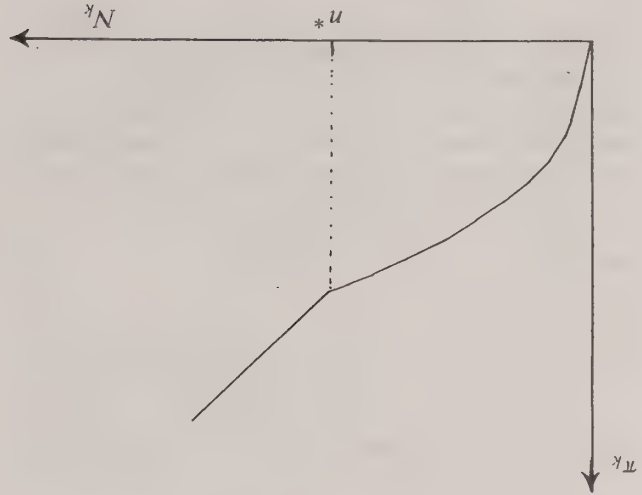


Figure 1. Graphique de π_k en fonction de N_k

2.2 Recherche d'un plan de sondage optimal

La variance maximum de P est fixée par les critères retenus pour le contrôle de qualité. Le sondage étant répété pour chacune des unités de traitement il est tout à fait naturel de chercher à minimiser l'espérance du coût du sondage donné en (2.1.1) soit:

(2.2.1)
$$E \sum^s (C_0 + n_k C_1) = \sum^U \pi_k (C_0 + n_k C_1).$$

Le problème d'optimisation s'écrit donc:

Minimiser
$$\sum^U \pi_k (C_0 + n_k C_1)$$

sous les contraintes:

$$N^{-2} \sum^U N_k^2 \left(\sigma^2 + \frac{n_k}{\tau^2} \right) \leq V_0$$

et pour tout k , $n_k \leq N_k$.

Associions un multiplicateur de Lagrange λ à la première contrainte – qui sera évidemment saturée – et des multiplicateurs μ_k aux autres. On obtient les solutions:

(2.2.2)
$$C_0 + n_k C_1 = \lambda \frac{N_k^2}{\tau^2} \left(\sigma^2 + \frac{n_k}{\tau^2} \right)$$

et, pour tout k :

(2.2.3)
$$C_1 \pi_k = \lambda \frac{N_k^2}{\tau^2} \cdot \frac{n_k}{\tau^2} + \mu_k$$

avec

$$\mu_k = 0 \text{ si } n_k < N_k \text{ et } \mu_k > 0 \text{ si } n_k = N_k.$$

Pour l'utilisation des multiplicateur de Lagrange, voir par exemple Luenberger (1973).

Pour toutes les unités primaires où $\mu_k = 0$ (les plus grosses) on obtient:

(2.2.4)
$$n_k = \frac{\sigma}{\tau} \left(\frac{C_1}{C_0} \right)^{1/2} = n^*.$$

Chaque unité primaire reçoit donc la même allocation, ce qui correspond à l'idée qu'on a besoin de la même précision dans chacune d'elle. Retournons à l'équation (2.2.3). On constate alors que, toujours pour ces unités primaires, les probabilités d'inclusion π_k doivent être proportionnelles aux tailles N_k soit:

(2.2.5)
$$\pi_k = \lambda^{1/2} C_1^{-1/2} \frac{\sigma}{\tau} N_k.$$

b)
$$E_{\xi} P_k (1 - P_k) = E_{\xi} (E_{\xi} ((P_k - P_k^2) | P_k))$$

$$= E_{\xi} P_k (1 - P_k) = \frac{N_k}{N_k - 1}$$

$$= (P(1 - P) - \sigma^2) \frac{N_k}{N_k - 1}$$

c) $E_{\xi} Z_k Z_{\ell} = 0$ à cause de l'indépendance des Z_k et des Z_{ℓ} , ce qui nous débarrasse d'un terme bien encombrant en même temps que des $\pi_{k\ell}$.

En récollant tous les morceaux de (2.3) et (2.4) un petit miracle algébrique se produit et nous avons l'expression:

$$E_{\xi} \text{Var } P \approx N^{-2} \sum^U N_k^2 \left(\sigma^2 + \frac{n_k}{\tau^2} \right)$$

(quantité positive par nature)

(2.1.1)
$$\cdot$$

avec $\tau^2 = P(1 - P) - \sigma^2$

Le miracle algébrique s'explique bien si on ne cherche pas à obtenir la variance sous le plan de sondage uniquement. Elle est d'ailleurs la conséquence d'un modèle un peu plus général que celui que nous avons posé.

Supposons que nous voulions estimer le total $N\bar{Y} = \sum^U Y_{\ell}$ d'une variable Y et que pour cela nous réalisions un tirage à deux degrés: un premier degré où des unités primaires k sont tirées avec des probabilités π_k , un second où n_k unités finales sont tirées par sondage aléatoire simple.

Nous posons un modèle où:

$$Y_{\ell} = Y + \alpha_k + \epsilon_{\ell},$$

avec α_k variable liée à l'U.P., d'indice k . Les α_k sont indépendantes de même loi d'espérance nulle de variance σ^2 . Les ϵ_{ℓ} sont également indépendantes centrées de variance égale à τ^2 . Avec $\pi_k^* = \pi_k n_k / N_k$ (taille de l'U.P. numéro k), l'estimateur de Horvitz-Thompson du total vaut $\bar{Y} = \sum Y_{\ell} / \pi_{\ell}^*$ la somme étant étendue à l'échantillon. Sous le modèle, et conditionnellement à l'échantillon on a:

$$\text{Var}_{\xi}(Y | s) = \sum^s \frac{\pi_k^2}{N_k^2} \left(\sigma^2 + \frac{n_k}{\tau^2} \right).$$

L'espérance sous le plan de cette expression redonne la formule (2.1.1).

2. ESTIMATION OPTIMALE DE LA PROPORTION

D'ENREGISTREMENTS ERRONNÉS À L'AIDE D'UN PLAN DE SONDAGE À DEUX DEGRÉS

Les unités primaires k (districts) comportent chacune un nombre connu N_k d'individus (des bulletins). Parmi ceux-ci D_k possèdent le caractère (être erronné). Le but est donc d'estimer :

$$P = \sum_k^U D_k / \sum_k^U N_k.$$

Le sondage consistera à tirer un échantillon s d'unités primaires (U.P.) avec des probabilités d'inclusion π_k au premier ordre et π_{kl} au second ordre à déterminer. Ensuite, si l'unité primaire k est tirée dans s on vérifiera n_k individus tirés par sondage aléatoire simple sans remise. Soit d_k le nombre de bulletins erronnés qu'on y relèvera.

L'estimateur P_k de $P_k = D_k/N_k$ sera $P_k = d_k/n_k$ et $\bar{D}_k = N_k P_k$ estimera D_k sans biais. L'estimateur de P sera :

$$\bar{P} = \frac{\sum_k^U D_k}{\sum_k^U \frac{\pi_k}{N_k}}. \quad (2.1)$$

C'est le ratio des estimateurs sans biais de D et de N , le nombre total de bulletins. Bien que ce nombre soit connu, il est bien évident que l'estimateur (4.1) est plus précis que $1/N \sum_k^U D_k / \pi_k$.

On a :

$$\text{Var}(\bar{P}) = E \text{Var}(\bar{P} | s) + \text{Var} E(\bar{P} | s). \quad (2.2)$$

Or :

$$\text{Var}(\bar{P} | s) = N^{-2} \sum_k^s N_k^2 \frac{\pi_k^2}{P_k(1-P_k)N_k} \left(\frac{n_k}{1} - \frac{N_k}{1} \right)$$

avec

$$N = \sum_k^s \frac{\pi_k}{N_k}.$$

D'où :

$$E \text{Var}(\bar{P} | s) = N^{-2} \sum_k^U \frac{\pi_k}{N_k^2} \frac{N_k(1-P_k)N_k}{N_k-1} \left(\frac{n_k}{1} - \frac{N_k}{1} \right). \quad (2.3)$$

Par ailleurs,

$$E(\bar{P} | s) = \frac{\sum_k^s \frac{\pi_k}{D_k}}{\sum_k^s \frac{\pi_k}{N_k}}.$$

La variance de cette quantité s'obtient par linéarisation en introduisant la variable $Z_k = D_k - P N_k = N_k(P_k - P)$.

On obtient :

$$\text{Var} E(\bar{P} | s) \approx N^{-2} \text{Var} \left(\sum_k^s \frac{\pi_k}{Z_k} \right).$$

Soit, compte tenu de ce que $\sum_k^U Z_k = 0$:

$$\text{Var} E(\bar{P} | s) = N^{-2} \left(\sum_k^k \frac{\pi_k}{Z_k^2} + \sum_{k \neq l}^k \sum_l \frac{\pi_k \pi_l}{Z_k Z_l} \pi_{kl} \right). \quad (2.4)$$

La somme des quantités (2.3) et (2.4) nous donne la variance de l'estimateur (2.1).

2.1 Introduction d'un modèle

La variance de \bar{P} est difficile à manipuler et, de plus, contient des paramètres inconnus. On se tire de la difficulté en faisant de nécessaires hypothèses qui se traduisent par un modèle de superpopulation. On supposera plus loin que les paramètres de ce modèle sont susceptibles d'être estimés à partir d'un essai préliminaire portant sur une toute petite partie de la population. On note E_k l'espérance sous le modèle (resp Var_k pour la variance) dont tous les aléas sont supposés indépendants du processus d'échantillonnage.

Le modèle suit les spécifications suivantes :

a) D_k suit une loi binomiale (N_k, p_k) . P_k est donc, sous le modèle, un estimateur de p_k .

b) p_k est lui même aléatoire. On suppose les p_k indépendantes et de même loi avec :

$$E_k p_k = P,$$

$$\text{Var}_k p_k = \sigma^2$$

pour tout k , quelle que soit, en particulier, la valeur de N_k .

En conditionnant, dans le modèle, par les p_k on a évidemment :

$$E_k(D_k | p_k) = N_k p_k,$$

$$\text{Var}_k(D_k | p_k) = N_k p_k(1 - p_k).$$

La variance anticipée de \bar{P} est la quantité $E_k \text{Var} \bar{P}$. C'est à elle que nous allons nous intéresser désormais. Pour l'évaluer on remarque que :

$$a) E_k(P_k - P)^2 = E_k(E_k(P_k - P)^2 | p_k) =$$

$$= \frac{N_k}{P(1-P) - \sigma^2} + \sigma^2,$$

Plans de sondage à deux degrés d'estimateurs de ratios: application au contrôle de qualité du recensement français de 1990

JEAN-CLAUDE DEVILLE¹

RÉSUMÉ

Cette étude est basée sur l'utilisation de modèles de superpopulation pour anticiper la variance d'une mesure par sondage de ratios *avant* l'enquête. On arrive, en utilisant des modèles simples qu'on voudrait néanmoins assez réalistes, à des expressions plus ou moins complexes qu'on parvient à optimiser, parfois rigoureusement, quelquefois de façon approximative. La solution du dernier des problèmes évoqués fait apparaître un facteur assez peu étudié en matière d'optimisation de plan de sondage: le coût lié à la mobilisation d'une information individuelle.

MOTS CLÉS: Contrôle de qualité du recensement; modèle de superpopulation; optimisation d'un plan à deux degrés; enquête à objectifs multiples.

1. INTRODUCTION

Le contrôle par sondage de la qualité des données du recensement Français a posé quelques problèmes à la fois intéressants et nouveaux. Trois d'entre eux sont traités dans cet article. Nous les étudierons en termes généraux et nous décrirons ensuite leur application au cas du recensement.

Dans tous les cas, le problème est celui de l'optimisation d'un sondage à deux degrés où les unités primaires sont des districts de collecte du recensement. Ceux-ci sont repérés par un indice k variant dans une population U de districts, qui est concrètement une unité de traitement des bulletins collectés.

Le premier problème consiste à estimer la fréquence d'un caractère dans la population de bulletins (le fait de comporter une erreur). Désirant avoir une précision donnée pour cette estimation on cherche à minimiser le coût du sondage avec une fonction de coût de la forme:

$$C_T = mC_0 + nC_1, \quad (1.1)$$

où m est le nombre d'unités primaires (districts) échantillonnées, C_0 le coût de traitement d'une U.P., n le nombre d'unités finales (bulletins) échantillonnées, et C_1 le coût de traitement d'une unité finale. Le problème est assez habituel dans le cas de l'estimation d'une moyenne (voir par exemple W. Cochran (1977)). Il reçoit ici une solution plus complète tenant compte de la grande variabilité de taille des unités primaires.

Le second problème est plus original et possède une portée plus générale. La population finale (ici les bulletins) est composée de G groupes ($g = 1$ à G) distincts. On

$$C_T = mC_0 + \sum_{g=1}^G n_g C_g, \quad (1.2)$$

désire avoir une estimation de la fréquence du caractère dans chacun des groupes, avec une précision donnée pour chacun d'eux. La contrainte réside dans le fait que les unités primaires seront communes à tous les groupes, l'échantillonnage au sein de chaque U.P. portant alors sur chaque groupe.

L'objectif est alors de minimiser le coût du sondage celui-ci ayant la forme:

où n_g est le nombre total d'unités finales du groupe g et C_g le coût de traitement d'une unité finale du groupe g . En pratique les groupes sont constitués par les différents types de bulletins utilisés dans le recensement.

Le troisième problème a sa source dans le contrôle de la codification. On possède, *a priori*, une mesure de la difficulté de codification de chaque bulletin. Formellement donc, on dispose au niveau de chaque individu i de la population, d'une variable quantitative X_i , telle que la probabilité – en un sens à préciser – pour que l'individu possède le caractère à mesurer soit à peu près proportionnelle à X_i . On cherche à utiliser cette information pour minimiser le coût du contrôle (mesure de la fréquence du caractère "codification erronée") sous la requête d'une précision donnée du sondage.

Dans chacun des cas, on utilise des modèles de superpopulation plausibles et simples qui permettent d'évaluer la variance anticipée du sondage. On a, en quelque sorte, une illustration presque typique d'un "échantillonnage assisté par un modèle" dans l'esprit du livre de Särndal, Swensson, Wretman (1992).

¹ Jean-Claude Deville, Chef de la Division des Méthodes Statistiques et Sondages, Institut National de la Statistique et des Etudes Economiques, 18, boul. Adolphe Pinard, 75675 Paris, CEDEX 14.

défini les expressions pour les valeurs infra-annuelles et annuelles ajustées et les covariances asymptotiques correspondantes. La méthode décrite dans cet article semble offrir un ajustement précis pour les données sur le commerce de détail au Canada. Cependant, il est nécessaire d'élaborer des tests de validité de l'ajustement pour ce modèle et pour d'autres modèles d'étalonnage.

REMERCIEMENTS

Nous tenons à remercier le rédacteur en chef, le rédacteur associé et deux arbitres pour leurs précieux conseils qui ont contribué à hausser grandement la qualité de cet article. Nous voulons aussi exprimer notre gratitude à messieurs J. Gambino et M. Kováčević, de Statistique Canada, pour les commentaires qu'ils ont faits sur cette étude.

BIBLIOGRAPHIE

BOX, G.E.P., et JENKINS, G.M. (1976). *Time Series Analysis, Forecasting and Control*. New York: Holden-Day.

CHOLETTE, P.A. (1984). L'ajustement des séries infra-annuelles aux repères annuels. *Techniques d'enquête*, 10, 39-53.

CHOLETTE, P.A. (1987). Benchmarking and interpolation of time series. Direction de la méthodologie, document de travail, Statistique Canada.

CHOLETTE, P.A. (1988). Benchmarking systems of socioeconomic time series. Direction de la méthodologie, document de travail, Statistique Canada.

CHOLETTE, P.A. (1992). Users' manual of programmes BENCH and CALEND to benchmark, interpolate and Calendarize time series data on micro computers. Direction de la méthodologie, document de travail, Statistique Canada.

CHOLETTE, P.A., et DAGUM, E.B. (1989). Benchmarking socio-economic time series data: a unified approach. Direction de la méthodologie, document de travail, Statistique Canada.

CHOLETTE, P.A., et DAGUM, E.B. (1991). Benchmarking time series with autocorrelated sampling errors. Direction de la méthodologie, document de travail, Statistique Canada.

DENTON, F.T. (1971). Adjustment on monthly or quarterly series to annual totals: An approach based on quadratic minimization. *Journal of the American Statistical Association*, 66, 99-102.

GODAMBE, V.P. (1960). An optimum property of regular maximum likelihood estimation. *Annals of Mathematical Statistics*, 13, 1208-1211.

HIDIROGLOU, M.A., et GIROUX, S. (1986). Composite estimation for the Retail Trade Survey. Direction de la méthodologie, document de travail, Statistique Canada.

HILLMER, S.C., et TRABELSI, A. (1987). Benchmarking of economic time series. *Journal of American Statistical Association*, 82, 1064-1071.

LANIER, N., et FYFE, K. (1989). Benchmarking of economic time series. Direction de la méthodologie, document de travail, Statistique Canada.

LANIER, N., et FYFE, K. (1990). Etalonnage des séries économiques. *Techniques d'enquête*, 16, 283-289.

LANIER, N., et MIAN, I.U.H. (1991). Maximum likelihood estimation for the constant bias model with mixed benchmarks. Direction de la méthodologie, document de travail, Statistique Canada.

MCLEOD, I. (1975). Derivation of the theoretical autocovariance function of autoregressive-moving average time series. *Applied Statistics*, 24, 255-256.

MIAN, I.U.H., et LANIER, N. (1991). Maximum likelihood estimation for the constant bias benchmarking model. Direction de la méthodologie, document de travail, Statistique Canada.

RAO, C.R. (1973). *Linear Statistical Inference and Its Applications*. (2^{ème} Ed.). New York: John Wiley.

Newton-Raphson a convergé très rapidement vers une solution. De fait, elle a convergé en 6 itérations seulement (environ 1 minute) – avec une précision de dix chiffres – tandis que la méthode des calculs successifs a convergé, pour le même degré de précision, au bout de 500 itérations et plus (plus de 45 minutes) sur un ordinateur personnel 386DX-25Mhz. Cependant, comme c'était prévu, les deux méthodes ont convergé vers la même solution finale. On obtient la matrice des covariances du vecteur estimé $(\hat{\theta}; \hat{\beta})'$ en inversant la matrice d'information de Fisher I , définie en (4.1), après avoir remplacé les paramètres par les estimations du MV correspondantes. Le tableau 3 donne la série originale des estimations mensuelles sur le commerce de détail de même que la série étalonée des estimations du MV avec les CV correspondants. On y trouve aussi la série des valeurs infra-annuelles ajustées et les CV

correspondants (deux dernières colonnes). La figure 1 représente graphiquement la série originale et la série étalonée. Ce graphique permet de constater que l'étalonnage ne modifie pas la tendance initiale de la série et qu'il s'opère une très forte réduction du CV des estimations infra-annuelles. Le tableau 4 donne la série originale des estimations annuelles sur le commerce de détail de même que la série des valeurs annuelles ajustées avec les CV correspondants. On calcule les variances des valeurs ajustées des tableaux 3 et 4 au moyen des expressions (4.4) et (4.5) respectivement, après avoir remplacé les paramètres par les estimations du MV correspondantes. Les résultats relatifs aux valeurs ajustées témoignent aussi d'une forte réduction du CV des estimations originales. Autrement dit, l'étalonnage a pour effet d'accroître la précision des séries mensuelles et annuelles.

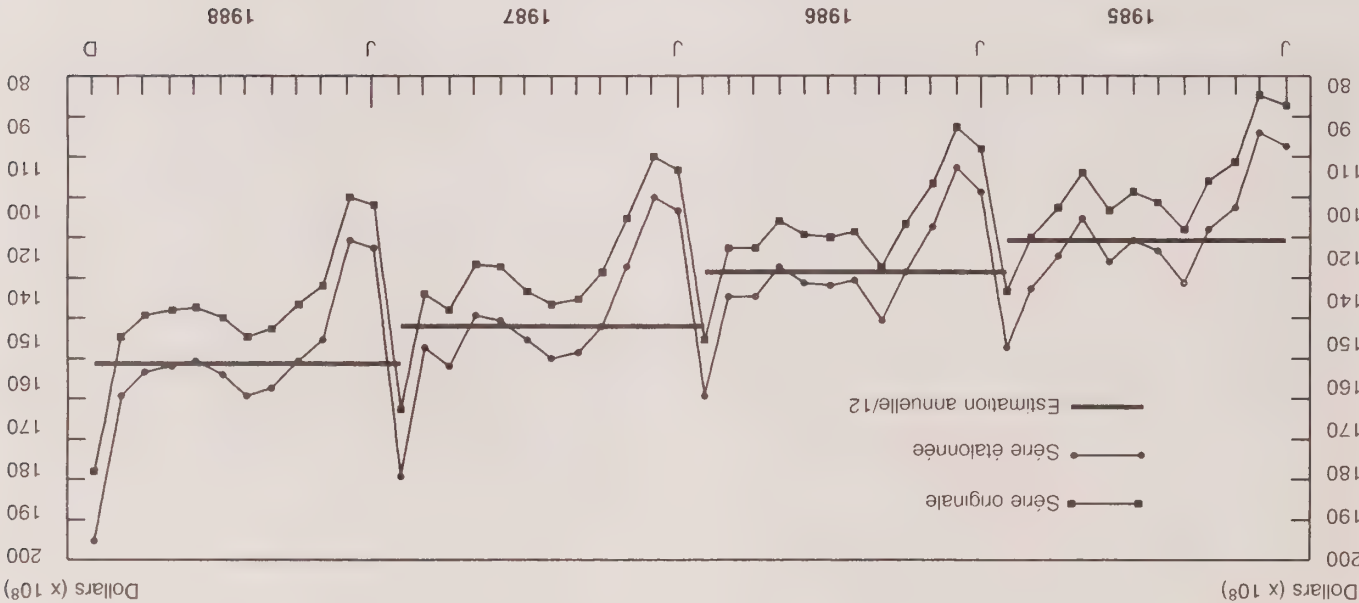


Figure 1. Estimations mensuelles sur le commerce de détail Canada, tous les magasins

Tableau 4

Estimations annuelles sur le commerce de détail et valeurs annuelles ajustées (en millions de dollars)

et CV correspondants

Année	z_T^*	CV (z_T)	z_T	CV (z_T)
1985	143,965,400	0.00033	143,927,507	0.00032
1986	154,377,100	0.00031	154,425,491	0.00030
1987	169,944,600	0.00193	169,101,697	0.00128
1988	181,594,000	0.00137	181,738,512	0.00127

*Source: Commerce de détail annuel, publication de Statistique Canada (n° 63-223 au catalogue, annuel).

7. CONCLUSIONS

Le modèle non linéaire traité dans cet article semble se prêter très bien à l'étalonnage de séries économiques construites à partir de grandes enquêtes par sondage. Les méthodes itératives proposées pour calculer les estimations du maximum de vraisemblance des paramètres du modèle sont très faciles à appliquer. Cependant, la convergence se fait beaucoup plus lentement avec la méthode des calculs successifs des équations d'estimation qu'avec la méthode de Fisher-Newton-Raphson. Nous avons présenté dans cet article les expressions en forme analytique fermée pour les covariances des estimateurs du MV. Ces estimateurs et les covariances correspondantes peuvent servir à faire des inférences sur les paramètres du modèle. Nous avons aussi

Tableau 3

Estimations mensuelles sur le commerce de détail, estimations du MV de Θ_t 's et des valeurs ajustées (en millions de dollars), et CV correspondants

Année Mois y_t^* $CV(y_t)^*$ Θ_t $CV(\Theta_t)$ \hat{y}_t $CV(\hat{y}_t)$

1985	1	8,689,668	0.008	9,686,630	0.00210	8,733,384	0.00667
	2	8,390,380	0.008	9,350,078	0.00210	8,429,951	0.00665
	3	10,107,485	0.006	11,248,048	0.00233	10,141,146	0.00496
	4	10,541,145	0.008	11,741,785	0.00200	10,586,294	0.00656
	5	11,763,659	0.007	13,094,151	0.00198	11,805,576	0.00570
	6	11,067,487	0.008	12,321,326	0.00189	11,108,803	0.00647
	7	10,810,755	0.008	12,029,467	0.00184	10,845,666	0.00643
	8	11,289,656	0.009	12,554,808	0.00206	11,319,309	0.00726
	9	10,336,540	0.009	11,484,216	0.00205	10,354,073	0.00728
	10	11,213,751	0.010	12,447,696	0.00256	11,222,737	0.00809
	11	11,935,495	0.010	13,234,412	0.00258	11,932,034	0.00808
	12	13,300,288	0.008	14,734,891	0.00188	13,284,853	0.00643

1986	1	9,753,373	0.009	10,794,009	0.00221	9,731,787	0.00716
	2	9,249,279	0.009	10,227,777	0.00224	9,221,277	0.00709
	3	10,609,952	0.008	11,729,293	0.00207	10,575,031	0.00622
	4	11,637,936	0.008	12,860,626	0.00206	11,595,032	0.00614
	5	12,695,108	0.008	14,024,139	0.00205	12,644,046	0.00605
	6	11,826,254	0.008	13,059,556	0.00202	11,774,385	0.00598
	7	11,940,908	0.010	13,164,500	0.00233	11,869,002	0.00740
	8	11,866,547	0.010	13,070,205	0.00232	11,783,987	0.00743
	9	11,540,397	0.009	12,712,283	0.00202	11,461,287	0.00670
	10	12,208,845	0.010	13,430,932	0.00235	12,109,215	0.00747
	11	12,201,498	0.010	13,418,219	0.00240	12,097,753	0.00747
	12	14,479,170	0.009	15,933,951	0.00215	14,365,916	0.00670

1987	1	10,271,723	0.012	11,276,676	0.00357	10,166,956	0.00891
	2	9,951,105	0.010	10,945,319	0.00261	9,868,208	0.00737
	3	11,492,162	0.008	12,663,849	0.00230	11,417,620	0.00584
	4	12,867,443	0.009	14,172,605	0.00235	12,777,901	0.00652
	5	13,508,434	0.012	14,850,145	0.00343	13,388,765	0.00862
	6	13,608,274	0.011	14,973,985	0.00287	13,500,418	0.00786
	7	13,278,474	0.023	14,483,340	0.01066	13,058,057	0.00165
	8	12,728,196	0.008	14,028,998	0.00227	12,648,426	0.00577
	9	12,616,239	0.009	13,888,982	0.00233	12,522,188	0.00659
	10	13,760,829	0.008	15,156,409	0.00227	13,664,890	0.00592
	11	13,380,142	0.008	14,733,240	0.00227	13,283,365	0.00597
	12	16,269,757	0.007	17,928,148	0.00241	16,163,867	0.00525

1988	1	11,134,013	0.010	12,234,529	0.00274	11,030,548	0.00753
	2	10,959,374	0.010	12,042,761	0.00276	10,857,651	0.00754
	3	13,177,788	0.008	14,508,565	0.00233	13,080,800	0.00602
	4	13,666,311	0.009	15,035,737	0.00243	13,556,094	0.00676
	5	14,267,530	0.006	15,742,039	0.00379	14,192,890	0.00448
	6	14,432,944	0.009	15,884,130	0.00240	14,320,997	0.00673
	7	13,960,825	0.009	15,363,957	0.00240	13,852,014	0.00673
	8	13,691,315	0.008	15,073,691	0.00233	13,590,312	0.00606
	9	13,773,109	0.008	15,159,075	0.00235	13,667,294	0.00613
	10	13,900,743	0.009	15,279,950	0.00255	13,776,282	0.00696
	11	14,453,461	0.009	15,884,279	0.00260	14,321,132	0.00700
	12	17,772,990	0.009	19,529,791	0.00267	17,607,895	0.00702

*Source: Commerce de détail, publication de Statistique Canada (n° 63-005 au catalogue, mensuel).

Or, les opérations normales de traitement des données d'enquête ne permettent pas de connaître les estimations des covariances; il faudrait une étude pour les obtenir. Par conséquent, pour les besoins de notre exemple, nous avons supposé que les covariances des estimations annuelles sur le commerce de détail sont nulles.

Tableau 2

Coefficients d'autocorrélation approximatifs révisés $\rho^*(k)$ pour les estimations mensuelles sur le commerce de détail (calculés pour 47 décalages)

Décalage $\rho^*(k)$	Décalage $\rho^*(k)$	Décalage $\rho^*(k)$
Décalage $\rho^*(k)$	Décalage $\rho^*(k)$	Décalage $\rho^*(k)$

0	1.0000	12	0.9602	24	0.8896	36	0.8100
1	0.9758	13	0.9345	25	0.8647	37	0.7869
2	0.9555	14	0.9126	26	0.8433	38	0.7669
3	0.9391	15	0.8943	27	0.8253	39	0.7501
4	0.9266	16	0.8798	28	0.8107	40	0.7363
5	0.9177	17	0.8687	29	0.7994	41	0.7254
6	0.9126	18	0.8612	30	0.7913	42	0.7176
7	0.9113	19	0.8572	31	0.7864	43	0.7126
8	0.9136	20	0.8567	32	0.7843	44	0.7106
9	0.9196	21	0.8595	33	0.7862	45	0.7114
10	0.9293	22	0.8661	34	0.7909	46	0.7151
11	0.9429	23	0.8760	35	0.7989	47	0.7217

Un des arbitres a soulevé une question intéressante: qu'arrive-t-il lorsque les variances et les covariances des estimations de l'enquête sont inconnues? C'est une question difficile, à laquelle on ne peut répondre sans une réflexion approfondie. Cependant, le modèle présenté ici suppose que ces variances et covariances sont connues. En règle générale, il suffit que les équations d'estimation utilisées pour calculer les estimations du MV soient des estimateurs consistants des variances et des covariances. Il est courant, en étalonnage, d'estimer ces variances et ces covariances à l'aide des données de l'enquête puisqu'on ne connaît jamais les valeurs théoriques (voir, par exemple, Hillmer et Trabelsi 1987).

Les calculs requis pour notre exemple sont exécutés par un algorithme rédigé dans le langage de programmation GAUSS pour micro-ordinateurs. L'estimation initiale de β pour le processus itératif, d'après (3.7), est $\beta_0 = 0.9162$. L'estimation initiale du vecteur de paramètres θ est calculée à l'aide de (3.4), après que l'on a remplacé β par β_0 . La méthode de Fisher-Newton-Raphson et la méthode d'itération successive, dont il a été question dans la section 3, servent toutes deux à calculer les estimations du MV des paramètres du modèle. L'estimation initiale finale de β se rapproche sensiblement de l'estimation initiale et β est égale à 0.9016, avec un CV de 0.0065. Il est intéressant de constater que, dans cet exemple, la méthode de Fisher-

entre les estimations mensuelles. Se fondant sur des données mensuelles sur le commerce de détail, Hidiroglou et Giroux (1986) ont calculé les valeurs estimées des coefficients d'autocorrélation aux décalages 1, 3, 6, 9 et 12 pour trois sortes de strates dans plusieurs provinces du Canada. La moyenne des estimations des coefficients d'autocorrélation entre les estimations mensuelles sur le commerce de détail pour les strates formées de la province de l'Ontario et du code 60 de la Classification type des industries (industries des aliments, boissons et médicaments) a servi d'approximation pour les coefficients d'autocorrélation entre les estimations mensuelles sur le commerce de détail. Ces coefficients (moyens) d'autocorrélation, c'est-à-dire $\rho(k)$, figurent dans le tableau 1.

Tableau 1

Coefficients d'autocorrélation approximatifs $\rho(k)$ pour les estimations mensuelles sur le commerce de détail

Décalage k	1	3	6	9	12
$\rho(k)$	0.970	0.940	0.918	0.914	0.962

La méthode des moindres carrés ordinaires ainsi qu'un algorithme de McLeod (1975) pour le calcul de coefficients d'autocorrélation théoriques pour des séries temporelles autorégressives à moyennes mobiles ont servi à réviser les coefficients d'autocorrélation observés. On a ajusté un modèle multiplicatif saisonnier ARMA (1,0)(1,0)¹² aux cinq coefficients observés en minimisant la somme des carrés des écarts entre le coefficient d'autocorrélation observé et le coefficient d'autocorrélation théorique. On a ensuite calculé les coefficients d'autocorrélation pour tous les autres décalages pertinents à l'aide des paramètres du modèle estimés et de l'algorithme de McLeod mentionné plus haut. Étant donné que le modèle ARMA est correct pour les coefficients d'autocorrélation théoriques, cette méthode produit une estimation constante de la fonction d'autocorrélation. Le tableau 2 donne les coefficients d'autocorrélation approximatifs révisés pour 47 décalages; ces coefficients ont servi à calculer approximativement les covariances des estimations mensuelles sur le commerce de détail par multiplication avec les écarts-types.

Au moment de l'étude des estimations annuelles sur le commerce de détail existaient uniquement pour les années 1985 à 1988. On peut trouver ces estimations dans la publication de Statistique Canada intitulée Commerce de détail annuel (n° 63-223 au catalogue, annuel). Les variances correspondantes ne sont pas publiées; on a donc dû les calculer à l'aide des données de l'enquête. Les covariances entre les estimations mensuelles et annuelles sont nulles parce que les échantillons des enquêtes mensuelle et annuelle sur le commerce de détail ont été tirés indépendamment l'un de l'autre. Comme les estimations annuelles sur le commerce de détail reposent sur des échantillons qui ne sont pas indépendants, leur covariance est non nulle.

5. ESTIMATION PAR LA MÉTHODE DU MAXIMUM DE VRAISEMBLANCE

LORSQU' $V_{ab} = 0$

Dans cette section, nous examinons l'estimation par le maximum de vraisemblance des paramètres du modèle dans le cas particulier où les vecteurs d'erreurs a et b sont non corrélés (c.-à-d., $\text{Cov}(a, b) = V_{ab}^b = V_{ab}^a = 0$). Cette situation est observée normalement dans les enquêtes par sondage lorsque les échantillons infra-annuels et annuels sont tirés de façon indépendante. On peut voir un changement dans les résultats des sections 3 et 4 en substituant $V_{ab}^b = V_{ab}^a = 0$ dans les équations. Par exemple, dans ce cas particulier, les estimateurs du MV de θ et de β , définis en (3.4) et en (3.5), se réduisent à

$$\theta^* \equiv \hat{\theta}(\beta) = (\beta^2 V_{aa}^{-1} + D' V_{bb}^{-1} D)^{-1}$$

$$(\beta V_{aa}^{-1} y + D' V_{bb}^{-1} z)$$

et

$$\hat{\beta}^* \equiv \hat{\beta}(\theta) = [\theta' V_{aa}^{-1} y] / [\theta' V_{aa}^{-1} \theta],$$

respectivement. On doit résoudre ces équations successivement pour obtenir les estimations voulues.

De la même manière, les éléments de la matrice d'information de Fisher se ramènent à

$$\eta_{11}^* = \beta^2 V_{aa}^{-1} + D' V_{bb}^{-1} D,$$

$$\eta_{22}^* = \theta' V_{aa}^{-1} \theta,$$

$$\eta_{12}^* = \eta_{21}^* = \beta V_{aa}^{-1} \theta.$$

6. APPLICATION

Dans cette section, nous présentons un exemple qui utilise des données publiées sur le commerce de détail au Canada; ces données viennent des enquêtes mensuelles et annuelles qu'effectue Statistique Canada sur le commerce de détail. Les estimations mensuelles sur le commerce de détail et les coefficients de variation (CV) correspondants se trouvent dans la publication de Statistique Canada intitulée Commerce de détail (n° 63-005 au catalogue, mensuel). Il existe deux catégories d'estimations mensuelles sur le commerce de détail, à savoir les estimations provisoires et les estimations révisées. Dans l'exemple qui suit, nous avons utilisé les estimations révisées mais non désaisonnalisées (brutes). Comme les CV des estimations révisées ne sont pas disponibles, nous nous sommes servis des CV des estimations provisoires pour calculer une approximation de la variance des estimations mensuelles révisées. Les données utilisées portent sur la période de janvier 1985 à décembre 1988. Une autre difficulté que comporte cet exemple est de déterminer les coefficients d'autocorrélation

$$\eta_{22} = -E \left[\frac{\partial^2 \ln(L)}{\partial \beta^2} \right] = \theta' V_{aa}^{-1} \theta$$

et

$$\eta_{12} = \eta_{21} = -E \left[\frac{\partial^2 \ln(L)}{\partial \theta \partial \beta} \right] = X_{\beta}' V^{-1} X_{\theta}.$$

Par conséquent, la matrice d'information de Fisher d'ordre $(n + 1) \times (n + 1)$ est définie

$$\eta = \begin{bmatrix} \eta_{11} & \eta_{12} \\ \eta_{21} & \eta_{22} \end{bmatrix}. \quad (4.1)$$

Si nous inversons η au moyen de l'algèbre de partition des matrices, nous avons

$$\text{Cov}(\theta) = \eta_{11.2}^{-1},$$

$$\text{Var}(\hat{\beta}) = \eta_{22.1}^{-1},$$

$$\text{Cov}(\hat{\theta}, \hat{\beta}) = -\eta_{11.2}^{-1} \eta_{12} \eta_{22.1}^{-1} \quad (4.2)$$

$$= -\eta_{11}^{-1} \eta_{12} \eta_{22.1}^{-1},$$

où

$$\eta_{11.2} = \eta_{11} - \eta_{12} \eta_{22}^{-1} \eta_{21},$$

$$\eta_{22.1} = \eta_{22} - \eta_{21} \eta_{11}^{-1} \eta_{12}. \quad (4.3)$$

Une fois que la matrice de covariances η^{-1} est obtenue, on peut calculer les covariances asymptotiques des valeurs infra-annuelles ajustées \hat{y} à l'aide de la méthode de delta (voir, par ex., Rao 1973). Soit Δ la matrice des dérivées partielles d'ordre un de y par rapport aux éléments de $(\theta'; \beta)'$. De toute évidence, la matrice $n \times (n + 1)$ est $\Delta = (\beta I_n; \theta)$. Si on utilise la méthode de delta, la matrice des covariances asymptotiques de \hat{y} est définie par l'expression

$$\text{Cov}(\hat{y}) = \Delta \eta^{-1} \Delta'. \quad (4.4)$$

En outre, la matrice des covariances des valeurs annuelles ajustées \hat{z} , d'après la théorie de la distribution normale centrée réduite à plusieurs variables, est définie par l'expression

$$\text{Cov}(\hat{z}) = D \eta_{11.2}^{-1} D', \quad (4.5)$$

où D et $\eta_{11.2}$ sont définies en (2.4) et en (4.3) respectivement.

logarithmique (3.1) ou, ce qui revient au même, en minimisant le terme quadratique \bar{Q} (3.2). Pour ce modèle en particulier, les estimateurs du MV et les estimateurs des MCG des paramètres du modèle sont les mêmes et nous n'en ferons la distinction que si le besoin s'en fait sentir. Si nous calculons les dérivées partielles d'ordre un de $\ln(L)$ par rapport à θ et à β respectivement et si nous posons le résultat égal à zéro, nous avons

$$\begin{aligned} \frac{\partial \ln(L)}{\partial \theta} &= X_{\theta}' V^{-1} (w - X_{\beta} \theta) = 0, \\ \frac{\partial \ln(L)}{\partial \beta} &= X_{\beta}' V^{-1} (w - X_{\beta} \theta) = 0, \end{aligned} \quad (3.3)$$

Puisque $E(w) = X_{\beta} \theta$ selon le modèle (2.3), les équations ci-dessus sont des équations d'estimation au sens de Godambe (1960) et elles sont non biaisées du point de vue de l'information. Il est intéressant de noter que $X_{\beta}' V^{-1}$ et $X_{\theta}' V^{-1}$ ne dépendent pas de w , de sorte que les équations (3.3) convergent vers zéro et ont, par conséquent, des racines constantes pourvu que $E(w) = X_{\beta} \theta$. Autrement dit, même si V dans les équations ci-dessus est remplacé par une de ses estimations constantes, les équations donneront des estimations constantes des vecteurs θ et β . Il convient aussi de noter que ces équations sont non linéaires par rapport aux paramètres à estimer et qu'on ne peut obtenir d'expression formelle pour les estimateurs de θ et de β . On peut donc recourir à une méthode itérative comme la méthode bien connue de Fisher-Newton-Raphson (aussi appelée méthode "de la fonction de caractérisation" de Fisher) pour calculer les estimations. La section 4 donne les éléments de la matrice d'information de Fisher espérée nécessaire à l'application de la méthode de Fisher-Newton-Raphson.

Une autre manière de calculer les EMV des paramètres du modèle est de résoudre successivement les équations d'estimation (3.3). Si nous résolvons la première expression de (3.3), l'estimateur de θ , comme fonction de β , est défini

$$\hat{\theta} \equiv \hat{\theta}(\beta) = (X_{\beta}' V^{-1} X_{\beta})^{-1} X_{\beta}' V^{-1} w. \quad (3.4)$$

De même, si nous résolvons la seconde expression de (3.3), l'estimateur de β , comme fonction de θ , est défini

$$\hat{\beta} \equiv \hat{\beta}(\theta) = [\theta' V_{aa}^{-1} (y - V_{ab} V_{bb}^{-1} (z - D\theta))] / [\theta' V_{ab}^{-1} \theta], \quad (3.5)$$

où

$$V_{aa.b} = V_{aa} - V_{ab} V_{bb}^{-1} V_{ba}.$$

On obtient les estimations du MV de θ et de β en calculant successivement les équations (3.4) et (3.5) jusqu'à convergence. L'avantage de cette méthode par rapport à la méthode de Fisher-Newton-Raphson est qu'elle est facile à appliquer.

Estimation initiale de θ et β

Afin d'établir une estimation initiale pour β , disons β_0 , réécrivons le modèle (2.3) comme suit:

$$w^* = X_{\theta}^* \beta + u^*,$$

où $w^* = ((Dy)'; (z - D\theta)'), X_{\theta}^* = ((D\theta)'; 0)'$ et $u^* = ((Da)'; b)'$. L'EMV de β est donc défini par l'expression

$$\hat{\beta} = [X_{\theta}^* (V^*)^{-1} w^*] / [X_{\theta}^* (V^*)^{-1} X_{\theta}^*], \quad (3.6)$$

où

$$V^* = \text{Cov}(u^*) = \begin{pmatrix} DV_{aa} D' & DV_{ab} \\ V_{ba} D' & V_{bb} \end{pmatrix}.$$

Si l'on tient compte de ce que $E(z) = D\theta$ et si on remplace $D\theta$ par z dans (3.6), on peut établir une estimation initiale pour β au moyen de la formule

$$\hat{\beta}_0 = \left[\begin{pmatrix} 0 \\ z \end{pmatrix}' (V^*)^{-1} w^* \right] / \left[\begin{pmatrix} 0 \\ z \end{pmatrix}' (V^*)^{-1} \begin{pmatrix} 0 \\ z \end{pmatrix} \right]$$

$$= [z' (D V_{aa.b} D')^{-1} D y] / [z' (D V_{aa.b} D')^{-1} z]. \quad (3.7)$$

On peut établir une estimation initiale pour β à l'aide de l'équation (3.4) en remplaçant β par β_0 .

4. COVARIANCES DES ESTIMATEURS

Cependant, la convergence est ordinairement très lente pour ce genre d'algorithme. Dans la section 6, nous comparerons ces deux méthodes dans le but de vérifier leur vitesse de convergence. Une fois que l'on connaît les estimations du MV des paramètres du modèle, on peut déterminer les valeurs infra-annuelles ajustées $\hat{y} = \hat{\beta} \theta$ et les valeurs annuelles ajustées $\hat{z} = D\hat{\theta}$.

Dans cette section, nous établissons les formules des covariances asymptotiques des estimateurs du MV des paramètres du MBMC en inversant la matrice d'information de Fisher, dénotée par Ω . Les covariances asymptotiques des valeurs annuelles et infra-annuelles ajustées sont calculées au moyen de la méthode delta. Considérons tout d'abord le calcul des covariances des estimateurs du MV de θ et β . Les éléments de Ω (c.-à-d. les espérances négatives des dérivées partielles d'ordre deux de $\ln(L)$) sont définis par les expressions

$$\Omega_{11} = -E \left[\frac{\partial^2 \ln(L)}{\partial \theta \partial \theta'} \right] = X_{\beta}' V^{-1} X_{\beta},$$

lequel non seulement le biais mais aussi les erreurs sont multiplicatives. L'auteur se sert de la théorie des MCG pour calculer les estimations des paramètres du modèle après avoir soumis ce modèle à une transformation logarithmique. Laniel et Mian (1991) ont défini un algorithme pour calculer les estimations du maximum de vraisemblance d'un modèle d'étalement à biais multiplicatif constant avec données-repères mixtes (c.-à-d. un mélange de repères fermes et de repères non fermes). Les données-repères fermes sont en l'occurrence des estimations tirées d'un recensement (c.-à-d. des estimations à variance nulle) tandis que les données-repères non fermes sont des estimations fondées sur un échantillon. L'hypothèse de l'existence d'un biais multiplicatif constant se vérifiera dans la pratique si la fréquence de mise à jour de la base de sondage est relativement stable, c'est-à-dire si le taux de sous-décomptement pour cette base varie très peu d'une année à l'autre. Cette hypothèse implique aussi que le rapport moyen des valeurs de la variable étudiée d'une période à l'autre est le même pour les entreprises qui sont incluses dans la base de sondage et celles qui ne le sont pas. La nature du biais rattaché aux estimations infra-annuelles peut varier d'une série temporelle à l'autre. Cholette et Dagum (1991) ont proposé une méthode d'étalement qui suppose l'existence d'un biais additif constant dans les estimations infra-annuelles.

L'objet de cet article est d'estimer les paramètres du modèle de Laniel et Fyfe par la méthode du maximum de vraisemblance (MV); les résultats s'inspirent du rapport de Mian et Laniel (1991). Le modèle en question est décrit dans la section suivante. Ensuite, nous étudions deux processus itératifs qui permettent de calculer les estimations du maximum de vraisemblance (EMV) des paramètres du modèle. Dans la section 4, nous donnons les expressions en forme analytique fermée pour les covariances asymptotiques des estimateurs des paramètres du modèle et des valeurs ajustées. À titre d'exemple, nous servons des données publiées par Statistique Canada sur le commerce de détail au Canada.

2. MODÈLE À BIAIS MULTIPLICATIF

CONSTANT (MBMC)

Afin de répondre aux exigences des enquêtes économiques en matière d'étalement, Laniel et Fyfe (1989, 1990) ont proposé le modèle à biais multiplicatif constant (MBMC) défini ci-dessous. Ce modèle suppose que les estimations infra-annuelles biaisées y_t répondent à la relation

$$y_t = \beta \theta_t + a_t, \quad t = 1, 2, \dots, n \quad (2.1)$$

et que les estimations annuelles non biaisées z_T répondent à la relation

$$z_T = \sum_{t \in T} \theta_t + b_T, \quad T = 1, 2, \dots, m, \quad (2.2)$$

où les indices t et T désignent respectivement les périodes infra-annuelles et les périodes annuelles, θ_t est le paramètre (inconnu) infra-annuel fixe, β est un paramètre (inconnu)

et

$$V = \begin{pmatrix} V_{aa} & V_{ab} \\ V_{ba} & V_{bb} \end{pmatrix}.$$

où

$$\ln(L) = -\frac{(n+m)}{2} \ln(2\pi) - \frac{1}{2} \ln |V| - \frac{1}{2} Q, \quad (3.1)$$

Suivant le MBMC, la fonction de vraisemblance logarithmique peut s'écrire

3. ESTIMATION PAR LA MÉTHODE DU MAXIMUM DE VRAISEMBLANCE

I_n étant une matrice unité d'ordre n , θ étant un vecteur ou une matrice nuls d'ordre approprié et d_{Tt} étant une fonction indicatrice égale à 1 pour $t \in T$ et égale à 0 dans le cas contraire. On suppose que les vecteurs d'erreurs d'échantillonnage a et b suivent une distribution normale multivariée telle que $a \sim MN(0, V_{aa})$ et $b \sim MN(0, V_{bb})$. De plus, dans le cas général, a et b sont corrélés, ce qui signifie que $\text{Cov}(a, b) = V_{ab} = V_{ba}' \neq 0$. Nous allons voir dans la section suivante que, pour ce modèle, l'estimateur du maximum de vraisemblance (MV) de θ et de β est identique à l'estimateur des moindres carrés généralisé (MCG) des mêmes paramètres. Par conséquent, l'hypothèse concernant la normalité de a et b n'est nécessaire que pour obtenir la matrice d'information de Fisher (et, donc, les variances) des EMV.

$$X_\beta = (\beta I_n : D)', \quad X_\theta = (\theta' : 0)', \quad X_D = (0' : D)', \quad w = (y' : z')', \quad u = (a' : b')', \quad D = (d_{Tt}'), \quad (2.4)$$

où

$$w = X_\beta \theta + X_D \theta + u, \quad (2.3)$$

Avant d'aller plus loin, définissons les vecteurs colonnes $y = (y_1, y_2, \dots, y_n)'$, $z = (z_1, z_2, \dots, z_m)'$, $a = (a_1, a_2, \dots, a_n)'$, $b = (b_1, b_2, \dots, b_m)'$, et $\theta = (\theta_1, \theta_2, \dots, \theta_n)'$. Le MBMC, défini par les équations (2.1) et (2.2), peut être réexprimé sous la forme

additives. Le modèle ci-dessus est un modèle de type hybride (mixte), de biais constant rattaché à y_t , et a_t et b_T sont les erreurs d'échantillonnage rattachées à y_t et à z_T respectivement. Dans lequel le biais est multiplicatif mais les erreurs,

On peut obtenir les estimations du maximum de vraisemblance des paramètres du modèle θ et β (si l'on suppose V connu) en maximisant la fonction de vraisemblance

Estimation d'un modèle d'étalement à biais multiplicatif constant par la méthode du maximum de vraisemblance et application

IJAZ U.H. MIAN et NORMAND LANIEL¹

RÉSUMÉ

Nous considérons le cas de l'estimation d'un modèle d'étalement non linéaire par la méthode du maximum de vraisemblance, tel que le présentent Laniel et Fyfe (1989; 1990). Ce modèle tient compte des biais et des erreurs d'échantillonnage rattachés à la série originale. Comme on ne peut exprimer les estimateurs du maximum de vraisemblance des paramètres du modèle sous une forme analytique fermée, nous examinons deux méthodes itératives permettant de calculer les estimations du maximum de vraisemblance. Nous donnons aussi les expressions en forme analytique fermée pour les variances et les covariances asymptotiques des séries étalonnées et des valeurs ajustées. Pour illustrer les méthodes, nous nous servons de données publiées sur le commerce de détail au Canada.

MOTS CLÉS: Autocorrélations; modèle à biais; moindres carrés généralisés; erreurs d'échantillonnage.

1. INTRODUCTION

Les méthodes d'étalement servent très souvent à améliorer les estimations tirées d'enquêtes infra-annuelles; on utilise à cette fin les estimations correspondantes tirées d'enquêtes annuelles et que l'on appelle des estimations répétées. Cette opération se traduit généralement par une réduction du biais et de la variance des estimations infra-annuelles. Par exemple, les estimations tirées des enquêtes annuelles sur le commerce de détail peuvent servir à améliorer les estimations mensuelles du commerce de détail. Les estimations infra-annuelles renferment souvent un biais à cause des lacunes de la base de sondage au point de vue de la couverture. Le sous-dénombrement est attribuable au fait que les nouvelles entreprises tardent à être incluses dans la base de sondage et à l'absence, dans cette base, des entreprises sans salariés. En outre, la variance des estimations infra-annuelles est souvent plus élevée que celle des estimations annuelles correspondantes et on observe des covariances d'échantillonnage pour les estimations de périodes différentes à cause du chevauchement des échantillons. En revanche, on peut supposer les estimations annuelles non biaisées parce que, dans la pratique, les bases de sondage correspondantes présentent peu de lacunes au point de vue de la couverture. On peut trouver des discussions détaillées sur l'étalement dans Laniel et Fyfe (1989; 1990), Cholette (1987; 1988) et d'autres ouvrages. La littérature statistique présente plusieurs méthodes d'étalement des séries chronologiques. Dans une perspective de minimisation quadratique, Denton (1971) propose plusieurs méthodes pour étalonner une série temporelle unique. Cholette (1984), à son tour, propose une variante de la méthode proportionnelle d'ordre un de Denton, par laquelle il supprime la condition de départ dans le but d'éviter les effets transitoires. Les hypothèses que posent

les auteurs ont très peu de chances d'être vérifiées avec la plupart des séries économiques. En effet, les modèles que ces auteurs élaboreront supposent que le biais rattaché aux estimations infra-annuelles suit une marche aléatoire et que les observations infra-annuelles et annuelles ne renferment aucune erreur d'échantillonnage. En règle générale, les estimations proviennent d'enquêtes par sondage et sont donc exposées à l'erreur d'échantillonnage. Hillmer et Trabelsi (1987) ont proposé une autre approche pour l'étalement, qui repose sur un modèle ARMMI (voir, par exemple, Box et Jenkins, 1976). Bien que cette approche tienne compte des covariances d'échantillonnage des estimations infra-annuelles et annuelles, elle fait abstraction du biais dans les estimations infra-annuelles. Cholette et Dagum (1989) ont modifié l'approche de Hillmer et Trabelsi en remplaçant le modèle ARMMI par un modèle d'"intervention". Cette variante permet la modélisation d'effets systématiques dans les séries temporelles mais elle présente néanmoins les mêmes lacunes que le modèle de Hillmer et Trabelsi (Laniel et Fyfe 1990). Afin de suppléer les lacunes mentionnées plus haut, Laniel et Fyfe (1989; 1990) ont proposé un modèle d'étalement non linéaire pour les mesures de niveau. Ils ont élaboré un algorithme complexe pour calculer les estimations par les moindres carrés généralisés (MCG) (et les covariances asymptotiques correspondantes) des paramètres du modèle. Ce modèle tient compte des covariances d'échantillonnage des estimations infra-annuelles et annuelles et peut être utilisé lorsque les données repères proviennent soit de recensements ou d'échantillons chevauchants annuels. Il suppose aussi l'existence d'un biais (relatif) multiplicatif constant dans les estimations infra-annuelles de niveau. D'autres modèles d'étalement à biais multiplicatif constant ont été proposés par Cholette (1992) et Laniel et Mian (1991). Cholette suppose un modèle dans

¹ Ijaz U.H. Mian et Normand Laniel, Division des méthodes d'enquêtes sociales, Statistique Canada, Ottawa, Ontario, K1A 0T6, Canada.

Le fait de laisser varier dans le temps les biais liés aux groupes de renouvellement est une extension naturelle du modèle, si l'on tient compte du fait que les moyennes des valeurs de la population varient en fonction du temps. Toutefois, la modélisation de l'évolution des biais liés aux groupes pourrait poser des difficultés, en raison d'éventuels problèmes d'identifiabilité touchant les modèles représentant la tendance et les effets saisonniers. Voir l'analyse présentée dans Pfeffermann (1991).

Les deux dernières extensions sont importantes et méritent d'être étudiées, mais notre expérience des données sur l'emploi nous porte à croire qu'elles auront un effet très minime sur les estimateurs du modèle.

REMERCIEMENTS

Les auteurs désirent remercier les arbitres pour leurs commentaires et leurs suggestions utiles. Les travaux relatifs à cette étude ont été réalisés pendant que le premier auteur se trouvait à Statistique Canada dans le cadre de son programme de bourse de recherche.

BIBLIOGRAPHIE

- ANSLEY, C.F., et KOHN, R. (1986). Predicted mean square error for state-space models with estimated parameters. *Biometrika*, 73, 467-473.
- BATTESE, G.E., HARTER, R.M., et FULLER, W.A. (1988). An error-components model for prediction of county crop areas using survey and satellite data. *Journal of the American Statistical Association*, 83, 28-36.
- CHANG, I., TIAO, G.C., et CHEN, C. (1988). Estimation of time series parameters in the presence of outliers. *Technometrics*, 30, 193-204.
- DAGUM, E.B. (1980). *La méthode de désaisonnalisation X-11 ARMI*. N° 12-564 au catalogue, Statistique Canada, Ottawa, Ontario, KIA 0T6.
- HAMILTON, J.D. (1986). A standard error for the estimated state vector of a state-space model. *Journal of Econometrics*, 387-397.
- HARRISON, P.J., et STEVENS, C.F. (1976). Bayesian forecasting (avec discussion). *Journal of the Royal Statistical Society, Série B*, 38, 205-247.
- HARVEY, A.C. (1984). A unified view of statistical forecasting procedures (avec discussion). *Journal of Forecasting*, 3, 245-275.
- HARVEY, A.C., et TODD, P.H.J. (1983). Forecasting economic time series with structural and Box-Jenkins models (avec discussion). *Journal of Business and Economic Statistics*, 1, 299-315.
- LEE, H. (1990). Estimation des coefficients de corrélation de panel pour l'Enquête sur la population active du Canada. *Techniques d'enquête*, 16, 297-306.
- MARAVALL, A. (1985). On structural time series models and the characterization of components. *Journal of Business and Economic Statistics*, 3, 350-355.
- MORRIS, N.D., et PFEFFERMANN, D. (1984). A Kalman filter approach to the forecasting of monthly time series affected by moving festivals. *Journal of Time Series*, 5, 255-268.
- PFEFFERMANN, D. (1991). Estimation and seasonal adjustment of population means using data from repeated surveys. *Journal of Business and Economic Statistics*, 9, 163-175.
- PFEFFERMANN, D., et BURCK, L. (1990). Estimation robuste pour petits domaines par la combinaison de données chronologiques et transversales. *Techniques d'enquête*, 16, 229-249.
- PFEFFERMANN, D., et BARNARD, C.M. (1991). Some new estimators for small-area means with application to the assessment of farmland values. *Journal of Business and Economic Statistics*, 9, 73-84.
- SINGH, M.P., DREW, J.D., GAMBINO, J.G., et MAYDA, F. (1990). Méthodologie de l'enquête sur la population active du Canada. N° 71-526 au catalogue, Statistique Canada, Ottawa, Ontario, KIA 0T6.
- TILLER, R.B. (1992). Time series modeling of sample survey data from the U.S. current population survey. *Journal of Official Statistics*, 8, 149-166.

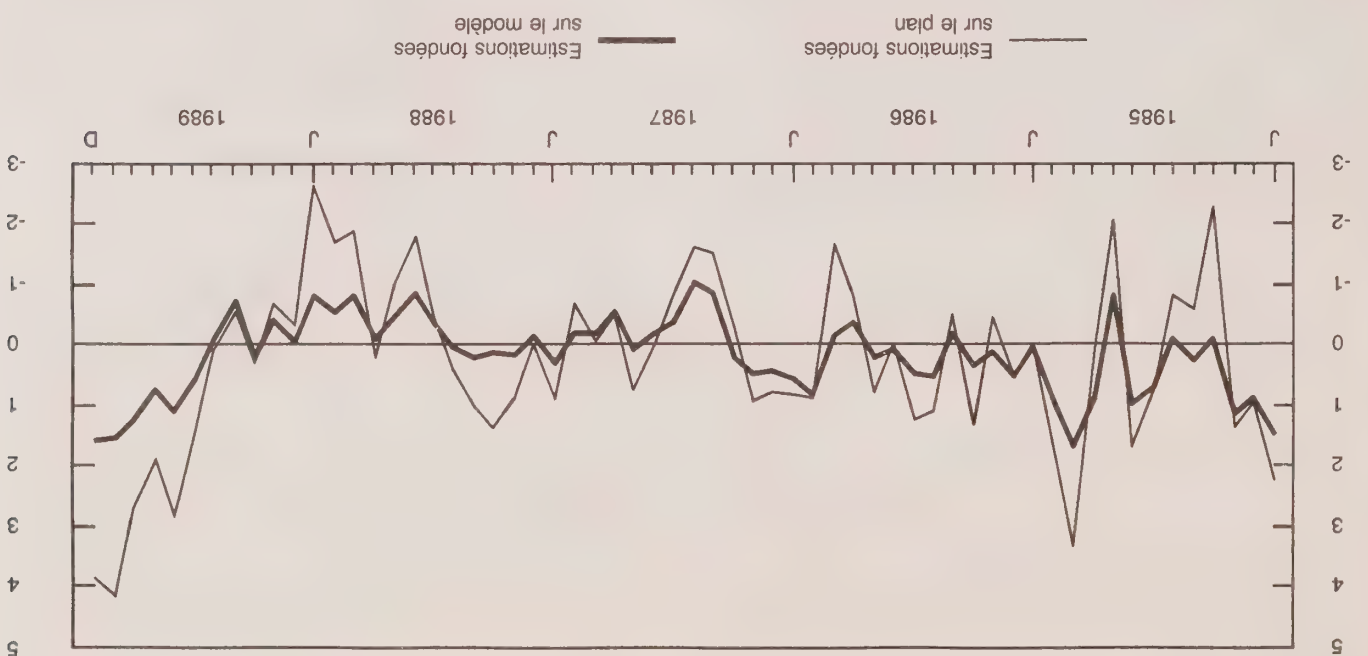


Figure 6. Variations annuelles des estimations fondées sur le plan et des estimations fondées sur le modèle de chômage de l'I.P.-E. ($\times 100$)

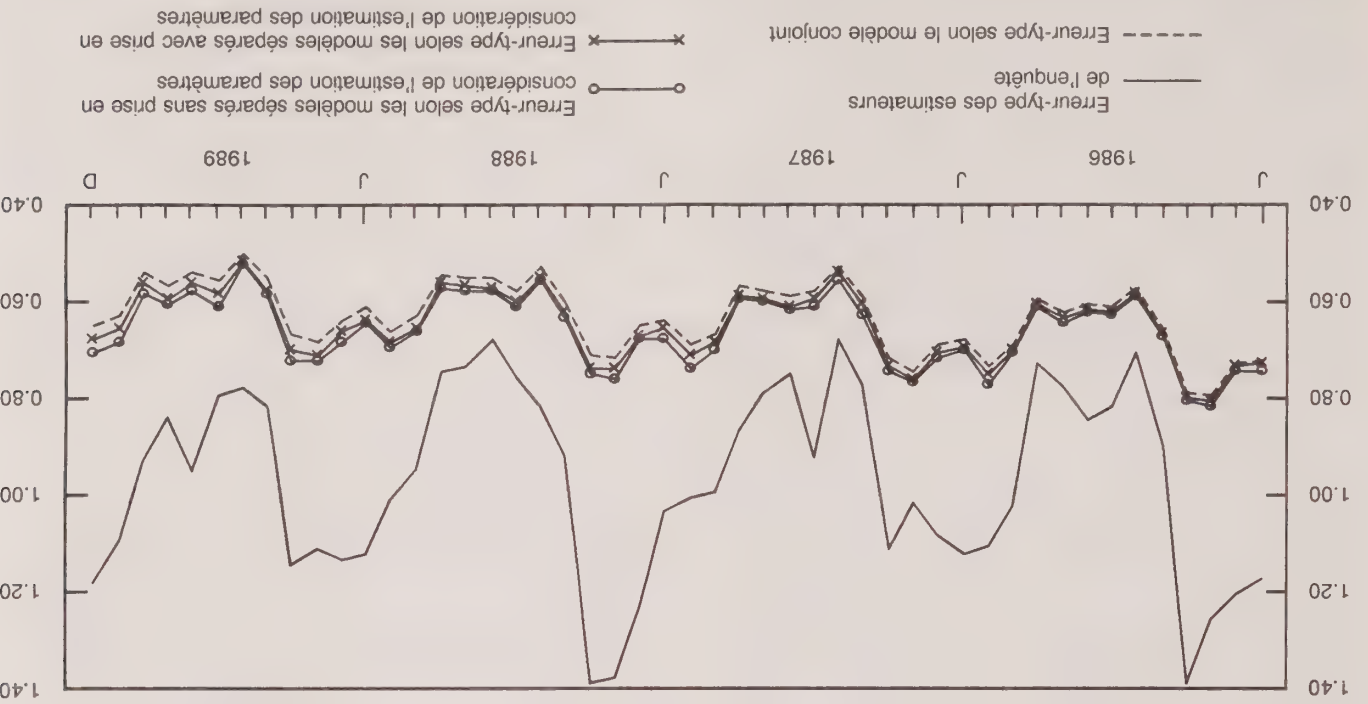


Figure 7. Erreurs-types des estimateurs de l'enquête et des estimateurs fondés sur le modèle avec et sans prise en considération des paramètres ($\times 100$) - Ile-du-Prince-Edouard

Comme on peut le voir, les effets saisonniers produits par les deux méthodes sont très voisins. Les estimations des niveaux de la tendance sont aussi voisines, mais la courbe de la tendance $X-1$ est plus lisse que la courbe du modèle. Une correspondance étroite semblable entre la procédure $X-11$ et le modèle est obtenue pour chacune des quatre provinces séparément, y compris, notamment, $I^1-P.-E.$, où les tailles des échantillons sont relativement faibles.

3.3 Comparaison entre les estimateurs fondés sur le plan et les estimateurs fondés sur le modèle

Il est mentionné, dans l'introduction, qu'une des principales raisons qui incitent à modéliser les estimateurs bruts de l'enquête est que le modèle produit des estimations des valeurs de la population qui, au moins pour les petites régions, sont plus exactes (lorsque le modèle tient) que celles de l'enquête. Nous avons calculé les deux ensembles d'estimations pour les quatre provinces et constaté que, comme prévu, les estimations produites par les deux méthodes se comportent de façon très semblable, mais que les estimateurs fondés sur le plan sont moins stables, affichant généralement des pointes et des creux plus prononcés. Un important aspect de la comparaison entre les deux ensembles d'estimations, c'est leur aptitude à estimer les variations annuelles des valeurs de la population. Une telle comparaison a l'avantage de ne pas être obscurcie par les effets saisonniers. La figure 6 montre les résultats obtenus pour $I^1-P.-E.$ Les estimations fondées sur le modèle sont les valeurs lissées du modèle conjoint qui utilisent toutes les données de tous les mois. Comme on peut le voir, les estimations produites par le modèle sont beaucoup plus stables et ne varient que légèrement d'un mois à l'autre, comparativement aux estimations fondées sur le plan. La figure 7 montre les erreurs-types ($E.-T.$) des estimateurs des taux de chômage pour $I^1-P.-E.$, calculées en vertu du plan (valeurs lissées, voir la figure 2) et en vertu du modèle conjoint. Sont également indiquées l'erreur-type résultant de l'ajustement du modèle séparé défini par (2.2), (2.5) et (2.6) et l'erreur-type correspondante après prise en considération du fait que des estimations des paramètres sont utilisées à la place des valeurs réelles inconnues. Voir la section 2.5 pour plus de détails. (Nous avons calculé cette dernière erreur-type uniquement pour le modèle séparé afin d'économiser du temps d'ordinateur.)

- 1) Les erreurs-types des estimateurs fondés sur le modèle, en vertu du modèle conjoint, ne sont que légèrement inférieures aux erreurs-types obtenues pour le modèle séparé, mais beaucoup plus faibles que celles des estimateurs de l'enquête.
- 2) Les erreurs-types des estimateurs fondés sur le modèle suivent le même profil que les erreurs-types des estimateurs de l'enquête, ce qui est une conséquence directe de la prise en considération, dans la définition du modèle, des changements des variances des erreurs de l'enquête au fil du temps. Voir la section 2.3 pour plus de détails.

graphiques:

Il y a trois aspects importants qui ressortent des

- 3) La prise en considération du fait qu'on utilise des valeurs estimées des paramètres dans le calcul de l'erreur-type des estimateurs fondés sur le modèle n'a qu'un effet minime sur l'erreur-type calculée. Rappe-lons que $I^1-P.-E.$ est la province ayant les plus faibles tailles d'échantillon. Pour les autres provinces, l'effet de la prise en considération de l'utilisation d'estimations des paramètres est encore plus faible.

4. RÉSUMÉ

Le présent article montre que des données recueillies en vertu d'un plan de sondage complexe, comprenant plusieurs degrés de sélection et des groupes de renouvellement, peuvent être modélisées avec succès par un modèle relatif-vement simple. Le modèle comprend deux parties: le modèle du recensement, qui représente les valeurs de la population, et le modèle des erreurs de l'enquête, qui décrit les relations qui existent dans les séries chronologiques des erreurs de l'enquête. L'utilisation du modèle donne des estimateurs plus exacts des valeurs de la population et de leurs composantes comme la tendance et la saisonnalité, et elle permet d'estimer l'erreur-type de ces estimateurs d'une façon relativement simple. On peut modifier les équations du modèle de manière à assurer la robustesse des estimateurs fondés sur le modèle et ainsi se protéger contre des défaillances possibles du modèle.

Le modèle utilisé dans cet article peut être prolongé dans diverses directions. Le premier effort devrait consister à appliquer simultanément le modèle à un plus grand nombre de provinces ou à d'autres petites régions, pour s'assurer que les estimateurs globaux de l'échantillon $\sum_{a=1}^A w_a y_a$ sont suffisamment proches des valeurs correspondantes de la population. Voir l'analyse présentée à la section 2.7. L'incorporation au modèle d'un mécanisme de détection des valeurs aberrantes, afin de mieux mesurer la performance et l'adéquation du modèle, est une autre addition valable.

Deux autres extensions pourraient consister à abandonner l'hypothèse d'une variance constante pour le terme d'erreur ϵ_i dans le modèle du recensement et à laisser varier dans le temps les biais liés aux groupes de renouvellement. La première extension découle de l'observation faite à la section 3.1 selon laquelle les variances des erreurs de l'enquête sont l'objet d'effets saisonniers, affichant un profil saisonnier semblable à celui des estimations brutes. L'ajustement des équations (2.4) dans les quatre provinces révèle également l'existence d'une légère tendance dans les variances qui, encore une fois, est semblable à celle des estimations brutes de l'enquête. Ainsi, les variances des erreurs de l'enquête semblent dépendre de la grandeur des estimateurs de l'enquête, ce qui laisse croire que les variances $\sigma_i^2 = V(\epsilon_i)$ changent avec le niveau des valeurs de la population. Comme première approximation, on pourrait supposer que σ_i^2 est proportionnel à la variance correspondante de l'erreur de l'enquête.

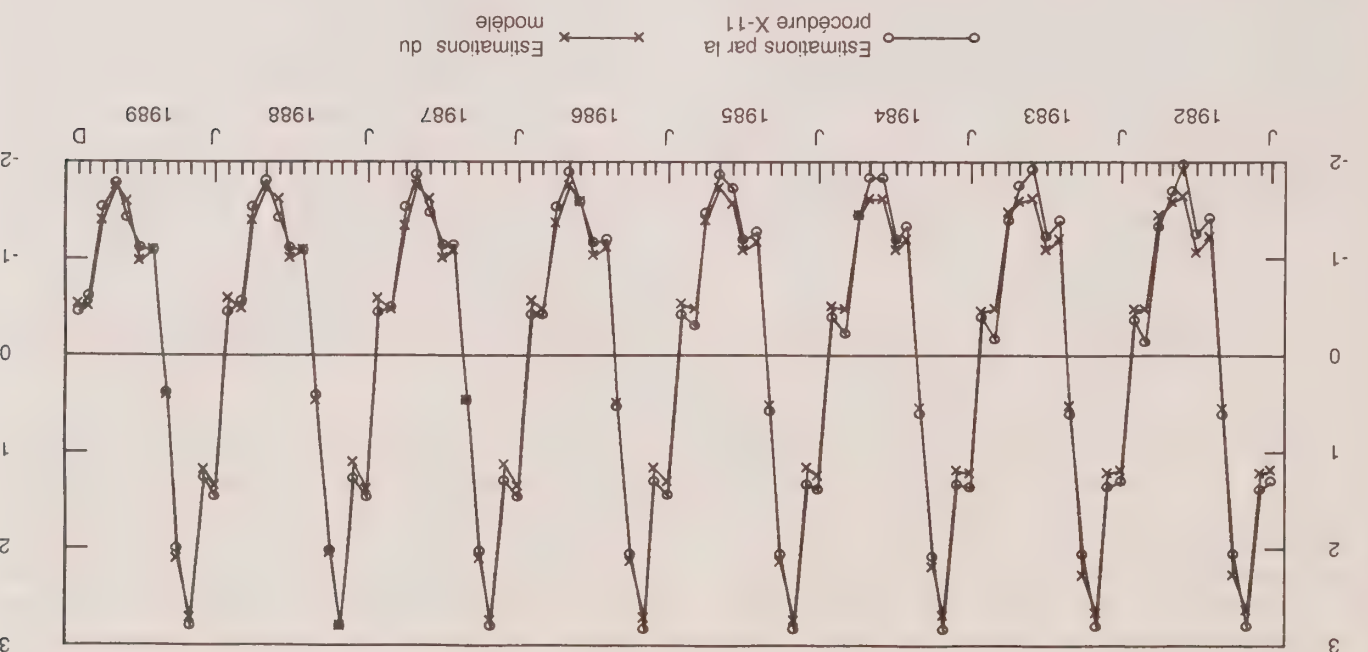


Figure 4. Moyennes pondérées des effets saisonniers par la procédure X-11 et selon le modèle ($\times 100$)

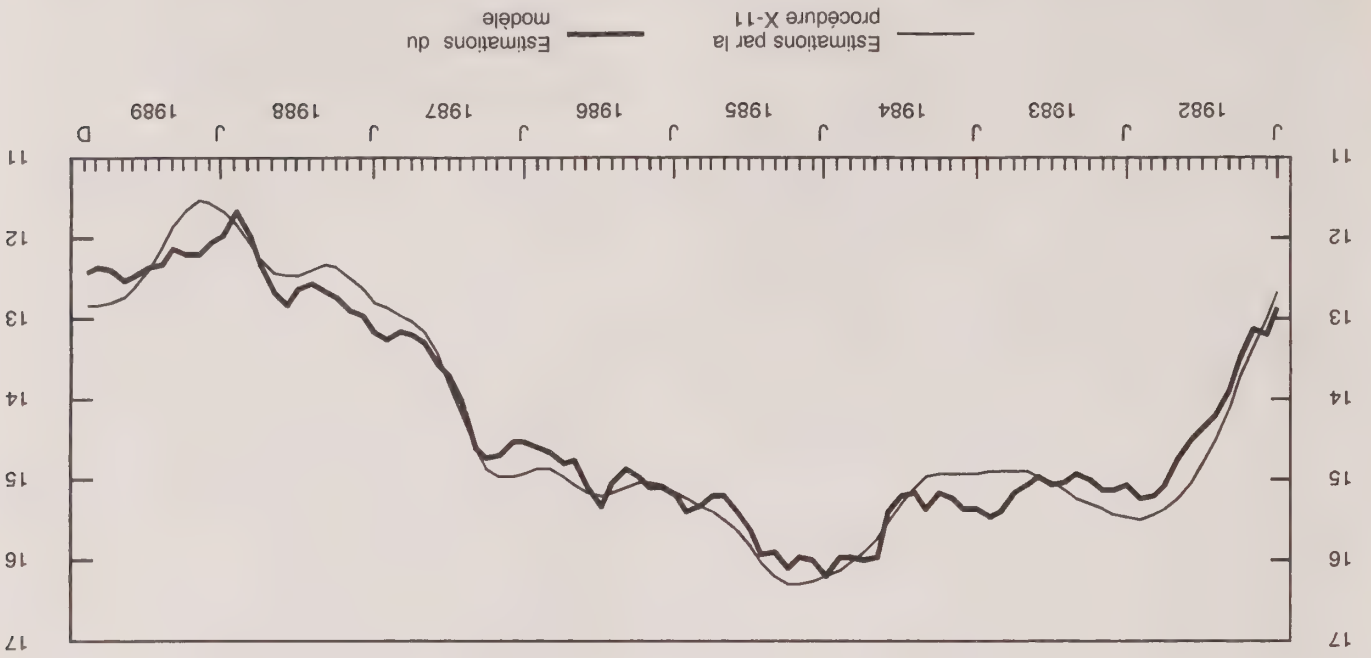


Figure 5. Moyennes pondérées des niveaux de la tendance obtenues par la procédure X-11 et selon le modèle

soudains, le modèle conjoint a une meilleure performance que les modèles séparés, même sans l'imposition des contraintes visant la robustesse. Par conséquent, en regroupant l'information des différentes provinces, le modèle conjoint s'adapte plus rapidement au nouveau niveau de la série. Pour d'autres illustrations de la performance des contraintes relatives à la robustesse, voir Pfeffermann et Burck (1990).

C. COMPARAISONS AVEC LES ESTIMATEURS PRODUITS PAR LA PROCÉDURE X-11

À titre d'évaluation finale du bien-fondé du modèle, nous comparons les estimations des effets saisonniers et des niveaux de la tendance produits par le modèle aux estimations produites par la procédure X-11 (Dagum 1980). Cette dernière est reconnue pour sa moins grande dépendance à l'égard des hypothèses particulières du modèle. Elle est la procédure couramment utilisée partout dans le monde pour la désaisonnalisation. La figure 4 présente les effets saisonniers moyens pour les quatre provinces obtenus avec la procédure X-11 et en vertu du modèle. La figure 5 montre les estimations correspondantes des niveaux de la tendance. Les moyennes sont calculées à l'aide des poids (w_{it}^a) employés dans les analyses précédentes. Les estimations fondées sur le modèle indiquées dans les deux figures sont les estimations lissées qui, comme celles de la procédure X-11, utilisent toutes les données de la période d'observation de l'échantillon.

Compte tenu des résultats très semblables obtenus pour les trois ensembles de prédicteurs examinés, et afin de faire ressortir la performance des contraintes relatives à la robustesse, nous avons délibérément appliqué une correction à la baisse de 33% aux taux de chômage de la période allant de mars 1985 à mars 1987, une correction à la baisse de 25% des taux de la période d'avril 1987 à novembre 1988 et une correction à la hausse de 33% des taux de la période de décembre 1988 à décembre 1989. Ces opérations ont pour effet d'introduire des mouvements soudains des données aux mois $t = 39$, $t = 64$ et $t = 84$. La figure 3 présente les erreurs de prédiction un mois d'avance globales (APE), les erreurs de prédiction un mois d'avance globales (APE), $I_t^a = \sum_{a=1}^4 w_{it}^a [\sum_{j=1}^6 (y_{it}^{(j)} - \hat{y}_{it}^{(j)}(t-1)) / 6]$, selon le modèle conjoint avec et sans les contraintes relatives à la robustesse, ainsi que selon les modèles séparés.

Il ressort clairement de la figure 3 que lorsque les contraintes sont imposées, les APE pour les périodes qui suivent les trois mois affichant des mouvements soudains sont moins élevées que les APE obtenues sans les contraintes. Par exemple, en mars 1985 ($t = 39$), les APE sont très élevées en valeur absolue avec et sans les contraintes, ce qui est évident étant donné que les prédicteurs ne sont fondés que sur les données allant jusqu'à février 1985. Les APE correspondant aux prédicteurs robustes reviennent toutefois à leur niveau normal beaucoup plus vite que les APE relatives aux prédicteurs non robustes. Une situation semblable peut être observée dans les deux autres périodes. Un autre résultat notable du graphique est que dans les périodes qui suivent les mois affichant des mouvements

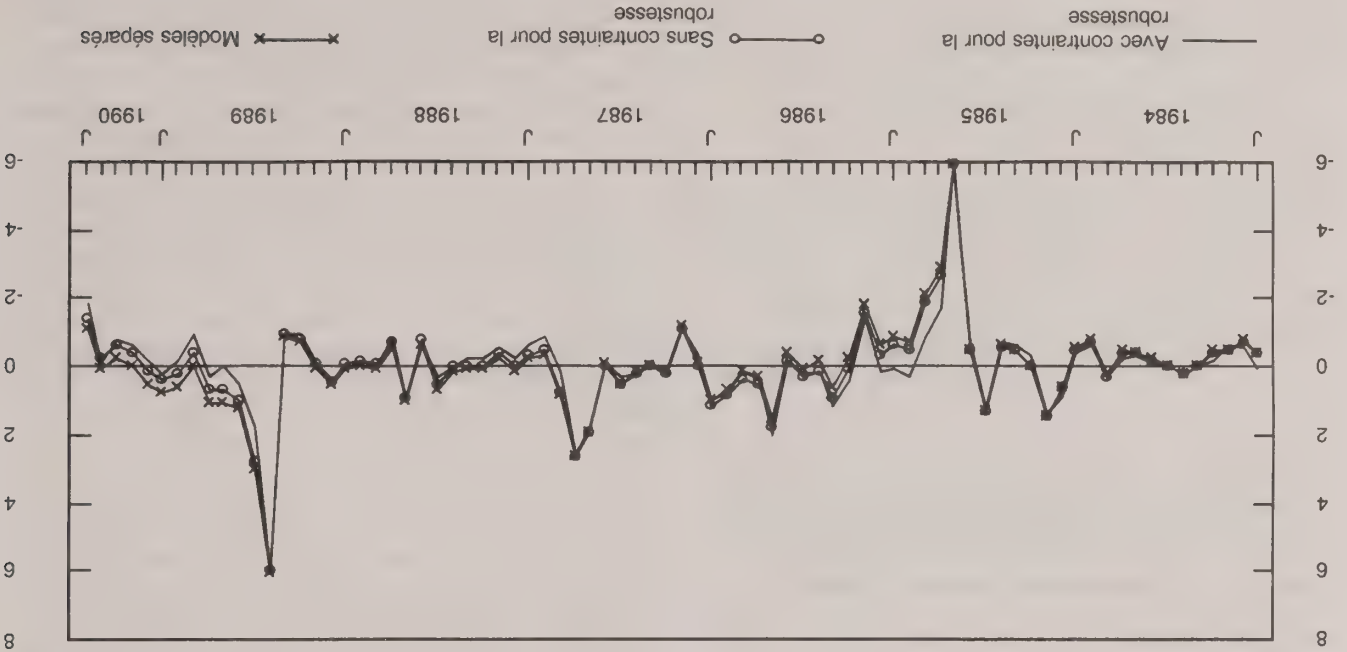


Figure 3. Erreurs de prédiction un mois d'avance globales pour les trois ensembles de prédicteurs (× 100) - Données contaminées

DIFFERENTS PRELECTEURS

Le tableau 3 présente des statistiques sommaires comparant le profil des erreurs de prédiction entre les quatre provinces, selon les résultats fournis par trois différents ensembles d'estimateurs des vecteurs d'états: 1) Les estimateurs obtenus en vertu des modèles séparés (MS) définis par (2.2), (2.5) et (2.6); 2) les estimateurs obtenus en vertu du modèle conjoint (MC) défini par (2.2), (2.5) et (2.10); 3) les estimateurs obtenus avec les contraintes imposées au modèle conjoint (ROB) pour les fins de la robustesse (2.11). Nous définissons ci-dessous les statistiques sommaires en utilisant encore une fois la notation $I_{(T)}^{(j)} = (Y_{(T)}^{(j)})^{(1)-(T-1)}$ pour l'erreur de prédiction obtenue lorsqu'on prédit l'estimateur du j -ième panel un mois à l'avance.

$$MB_a = \sum_{l=k+1}^N I_0^{(l)}(6)/(N-k) - \text{biais moyen de la pr\'ediction de l'estimateur moyen de l'enqu\^ete } \bar{y}_{ia} = \sum_{j=1}^6 y_{ij}^{ia}/6.$$

$$MAB_n = \sum_{j=1}^6 I_{k+1}^{(j)} / (N - k) \mid \sum_{j=1}^6 \text{absolu moyen de la pr\'ediction des estimateurs propres aux panels.}$$

$$S\tilde{Q}R E_a = \{ \sum_{i=1}^{k+1} [1/6 \sum_{j=1}^6 I_{ij}^{(1)} / y_{ij}^{(1)}] / \sum_{i=1}^k [1/6 \sum_{j=1}^6 I_{ij}^{(1)} / y_{ij}^{(1)}] / (N-k) \}$$

carrière moyenne quand on prédit l'estimateur moyen de l'enquête.

Les statistiques sommaires ci-dessus sont présentées séparément pour la période d'observation de l'échantillon allant de juillet 1983 à décembre 1988, et pour la période postérieure allant de janvier 1989 à décembre 1989. Dans ce dernier cas, les données ont été ajoutées un mois à la fois, c'est-à-dire que pour prédire l'estimateur de l'enquête de février 1989, par exemple, nous avons utilisé les données observées jusqu'à janvier 1989, et ainsi de suite.

Les principales conclusions qui découlent du tableau 3 sont les suivantes:

sont les suivantes:

Tableau 3

Erreurs de prédiction pour les quatre provinces – statistiques sommaires ($\times 100$)[illegible]

(7) Les résultats obtenus dans le cas des trois ensembles de prédicteurs sont en général très semblables, ce qui indique que pour les données analysées, l'utilisation du modèle conjoint ne produit qu'une légère amélioration par rapport à l'utilisation des modèles séparés, et qu'il n'y a pas de variations brusques du niveau des séries au cours des années examinées.

(2) Les erreurs de prédiction des estimateurs de l'enquête sont faibles tant à l'intérieur qu'à l'extérieur de la période d'observation de l'échantillon, ce qui laisse supposer un bon ajustement du modèle. Notons que sauf pour l'I.-P.-E., les erreurs de prédiction relatives mesurées par la statistique $SQRE_q$ sont toutes inférieures à 70%.

(3) Les biais des erreurs de prédiction dans la période postérieure à l'observation de l'échantillon sont plus élevés que dans la période d'observation de l'échantillon, les écarts étant particulièrement prononcés au Nouveau-Brunswick et à l'I.-P.-E. Ce résultat à lui seul pourrait laisser croire à une certaine défaillance du modèle au cours de l'année 1989. Toutefois, l'examen des erreurs de prédiction mensuelles des panels dans les quatre provinces pour cette année-là (non présentées dans le tableau) indique que même si les erreurs sont en général positives, les biais relativement élevés résultent principalement d'une ou deux erreurs extrêmes, qui ont un effet important sur les statistiques sommaires moyennes, puisqu'il n'y a que 12 mois de données. Il est à signaler, en outre, que les taux de chômage estimés dans les quatre provinces au cours de l'année 1989 se situent entre 0,11 et 0,18, de telle sorte qu'un biais de prédiction de .005, ou même de .009 comme dans le cas de l'I.-P.-E., n'est pas élevé. De toute évidence, le modèle peut être modifié pour tenir compte de ces biais, si ceux-ci persistent avec l'ajout de nouvelles données. Par ailleurs, rappelons que le biais des erreurs de prédiction, puisque le biais des estimateurs fondés sur le modèle des valeurs de la population correspondantes est régi par les contraintes imposées pour assurer la robustesse (2,11).

Comme il a été mentionné à la section 2.3, plutôt que d'utiliser les estimations originales des erreurs-types dues au plan dans les modèles ajustés aux erreurs propres aux panels, nous utilisons des valeurs lissées, ce qui réduit les effets des erreurs d'échantillonnage sur les estimateurs de l'enquête. La figure 2 présente le graphique des deux ensembles d'estimateurs pour l'Île-du-Prince-Édouard (I.-P.-É.), qui est la plus petite province de la région de l'Atlantique, et donc celle où les tailles des échantillons sont les plus faibles. Comme on peut le voir, l'effet du lissage est d'éliminer les estimations brutes extrêmes, mais, par ailleurs, les valeurs lissées suivent en gros le même profil que les estimations brutes. Les graphiques des autres provinces présentent des profils semblables, mais les écarts entre les estimations brutes et les estimations lissées sont moins prononcés en raison du fait que les tailles des échantillons sont plus élevées dans ces provinces.

Nous concluons cette section en définissant les modèles postulés pour les effets saisonniers dans les quatre provinces. Notre modèle initial supposait des variances fixes pour les termes d'erreur $\eta_{st} = \sum_{j=0}^J S_{t+j}$, $t = 1, 2, \dots$ (voir l'équation 2.1). Les erreurs prédites $\hat{\eta}_{st} = \sum_{j=0}^J \hat{S}_{t+j}$ fournies par ce modèle diminuaient en valeur absolue en fonction du temps dans trois des quatre provinces, et augmentaient dans la dernière. Notons qu'en vertu du modèle défini par (2.1), avec des variances constantes des termes d'erreur des états, le filtre de Kalman converge vers un état stable dans lequel les matrices V-C des estimateurs des vecteurs d'états, et donc les $\hat{\eta}_{st}$, sont constants. Donc, nous avons modifié le modèle initial de telle façon que $\text{VAR}(\eta_{st}) = \sigma_s^2 \times g(t)$ où, pour la Nouvelle-Écosse, Terre-Neuve et l'I.-P.-É., $g(t) = t^{(-3/2)}$, tandis que pour le Nouveau-Brunswick, $g(t) = t^{1/2}$.

3.2 Résultats

3.2.1 Biases liés aux groupes de renouvellement

Le tableau 2 montre les biais liés aux groupes de renouvellement (BGR) et leurs erreurs-types estimées (E.-T.) pour les quatre provinces, en vertu du modèle complet défini par (2.3), (2.5), (2.6) et (2.10).

Tableau 2

Biases liés aux groupes de renouvellement et erreurs-types pour les quatre provinces (× 100)

Panels	Nouvelle-Écosse		Nouveau-Brunswick		Terre-Neuve		I.-P.-É.	
	BGR	E.-T.	BGR	E.-T.	BGR	E.-T.	BGR	E.-T.
1	-0.20	0.10	-0.02	0.11	-0.47	0.13	0.32	0.17
2	0.18	0.09	0.40	0.10	0.42	0.12	0.18	0.15
3	0.32	0.08	0.24	0.09	0.47	0.12	0.31	0.15
4	0.06	0.07	0.01	0.09	0.18	0.12	0.03	0.15
5	-0.03	0.08	-0.15	0.10	-0.10	0.13	-0.25	0.16
6	-0.34	0.08	-0.50	0.11	-0.50	0.14	-0.60	0.16

Les biais liés aux groupes de renouvellement affichent un profil relativement uniforme entre les provinces. Ainsi, les biais relatifs au 3^e et au 6^e panel sont tous hautement significatifs selon la statistique t classique, affichant un signe positif pour le 3^e panel et un signe négatif pour le 6^e panel. Les biais relatifs au 4^e et au 5^e panel ont eux aussi le même signe dans toutes les provinces, et sont tous non significatifs.

Pour le 2^e panel, tous les biais sont positifs, mais celui de l'I.-P.-É. est non significatif (l'I.-P.-É. est la province ayant la taille d'échantillon la plus faible). C'est également pour l'I.-P.-É. que le signe du biais du 1^{er} panel est différent de celui des autres provinces.

Comme il a été indiqué à la section 2.3, il y a plus d'une raison possible pour expliquer l'existence de biais liés aux groupes de renouvellement, mais les résultats présentés dans le tableau incitent fortement à penser que quelle que soit la raison, les biais observés pour certains des panels sont réels, et non pas attribuables uniquement aux erreurs d'échantillonnage. Un inconvénient de la présente analyse, toutefois, est que les biais liés aux groupes de renouvellement sont supposés fixes dans le temps. Un modèle plus souple est proposé à la section 4.

3.2.2 Qualité de l'ajustement

A. TEST DE LA NORMALITÉ

Définissons par $I_t^{(j)} = (y_t^{(j)} - y_t^{(j)(t-1)})$ l'erreur de prédiction qu'on obtient en prédisant l'estimateur du j -ième panel un mois à l'avance, et posons $I_t^{(1)} = (I_t^{(1)})$, \dots , $I_t^{(6)}$. L'utilisation de l'estimation du maximum de vraisemblance dans cette étude suppose que les vecteurs $I_t^{(j)}$ sont des écarts aléatoires normaux (voir la section 2.4). Pour tester cette hypothèse, nous avons calculé la distribution empirique des erreurs de prédiction standardisées $\{(SI_t^{(j)}) / SD(I_t^{(j)})\}$, $t = (k + 1), \dots, N$, et nous l'avons comparée à la distribution normale réduite au moyen de la statistique de test de Kolmogorov-Smirnov. La statistique de test a été calculée pour chacun des six panels et chacune des quatre provinces, et a donné des valeurs P supérieures à 0.15 dans 21 cas sur 24. (Les tests ont été faits au moyen de la procédure PROC UNIVARIATE du logiciel SAS. Dans cette procédure, si la taille de l'échantillon est plus grande que cinquante, comme dans notre cas, les données sont testées à l'égard d'une distribution normale avec moyenne et variance égales à la moyenne et à la variance de l'échantillon.) L'application de la même procédure de test aux erreurs de prédiction standardisées $\{(SI_t^{(j)}) / SD(I_t^{(j)})\}$, $t = (k + 1), \dots, N$, où $I_t^{(j)} = [\sum_{j=1}^6 I_t^{(j)}] / 6$, donne des valeurs P supérieures à 0.15 dans chacune des quatre provinces.

Les estimateurs des écarts-types des erreurs de prédiction ayant servi aux tests sont ceux produits par le filtre de Kalman, sans prise en considération de la composante de variance résultant de l'estimation des paramètres (voir la section 2.5). Cette dernière composante est négligeable même pour l'I.-P.-É., qui compte les échantillons ayant la taille la plus faible parmi les quatre provinces. Nous reviendrons sur cette observation à la section 3.4.

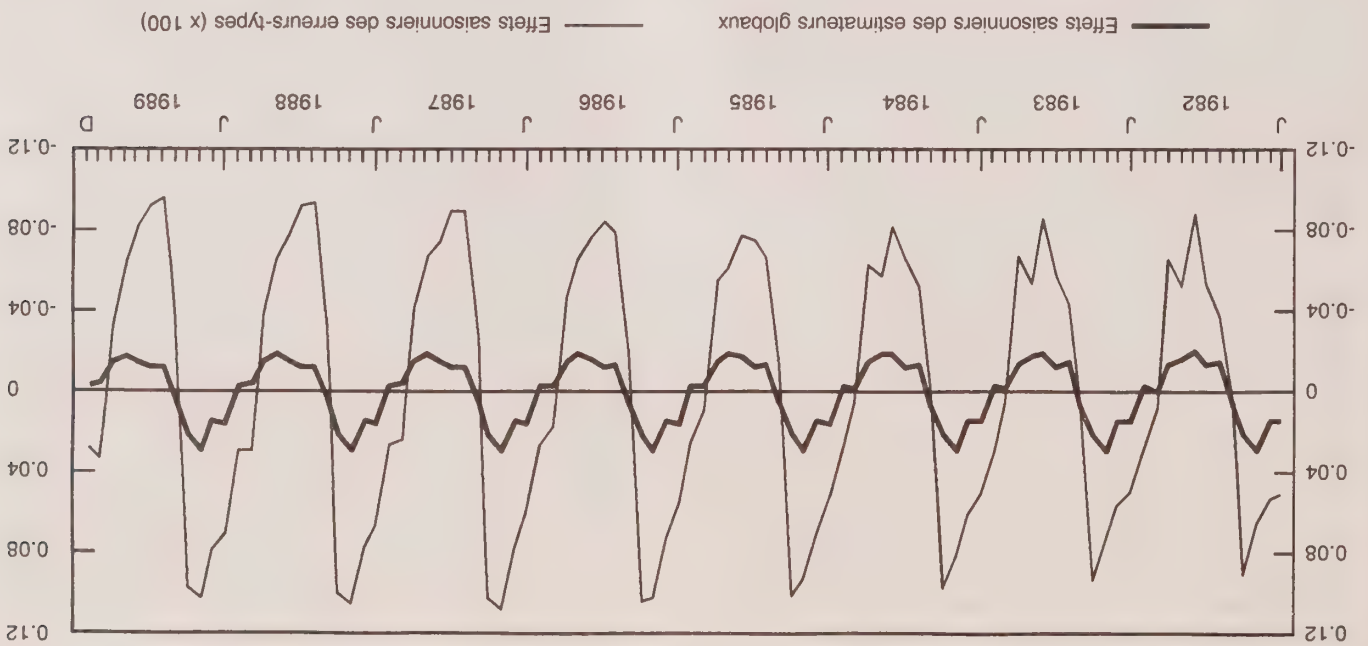


Figure 1. Effets saisonniers des estimateurs globaux de l'enquête et des erreurs-types des estimateurs globaux de l'enquête ($\times 100$)

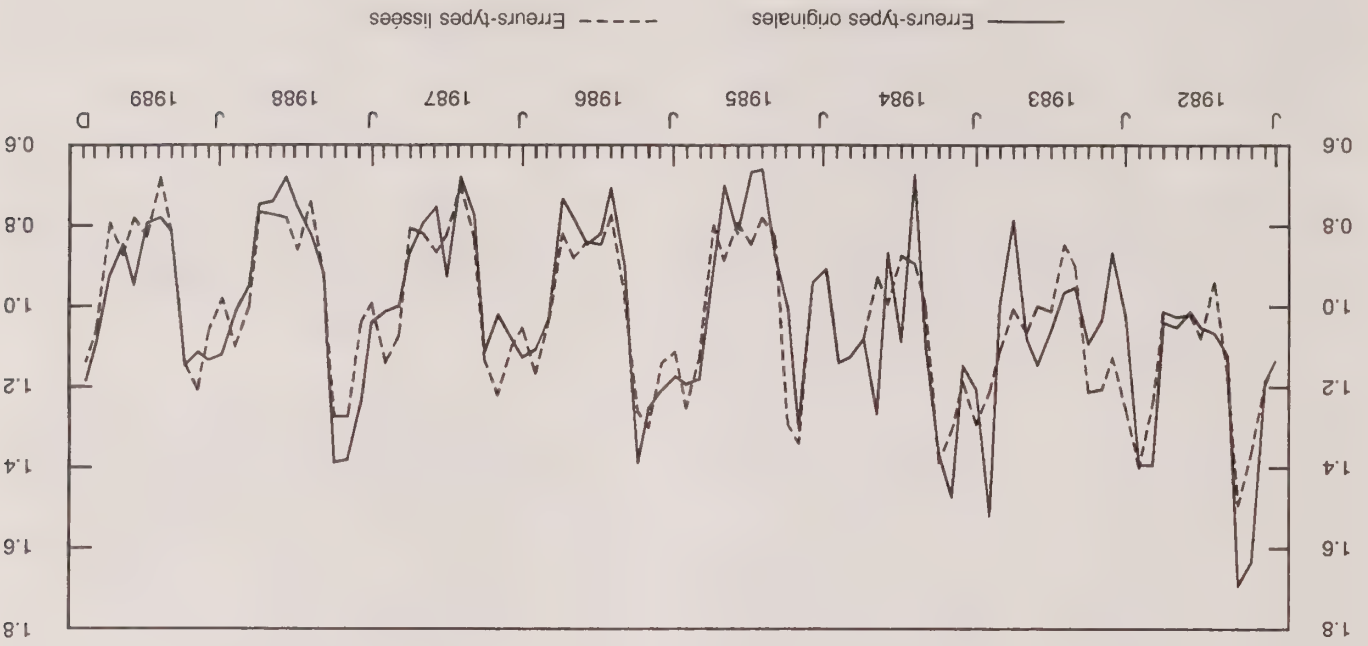


Figure 2. Erreurs-types originales et lissées des estimateurs de l'enquête ($\times 100$) - Île-du-Prince-Édouard

3. AJUSTEMENT DU MODÈLE AUX PROVINCES

DE L'ATLANTIQUE:
RÉSULTATS EMPIRIQUES

Le modèle défini par (2.2), (2.5), (2.6) et (2.10) a été ajusté aux estimateurs de panel mensuels pour les quatre provinces de l'Atlantique, en deux étapes. À la première étape, le modèle défini par (2.2), (2.5) et (2.6) a été ajusté à chacune des provinces séparément. À la deuxième étape, les corrélations définissant la matrice ϕ de (2.10) ont été estimées au moyen d'une recherche par quadrillage (voir la section 2.6). Les estimateurs obtenus sont $\text{Diag}(\phi) = (0.5, 0.25, 0.80, 0.0)$. Les données ayant servi à l'estimation du modèle portent sur les années 1982 à 1988. Les données de 1989 ont servi à faire le diagnostic du modèle, en comparant les résultats obtenus à l'intérieur et à l'extérieur de la période d'observation de l'échantillon.

3.1 Analyse préliminaire

Le tableau 1 présente une ventilation de la population active des quatre provinces, par industrie. Les chiffres du tableau reflètent la situation qui existait en mars 1991. Les tailles (attendues) des échantillons de l'EPA sont également indiquées. Comme on peut le voir, les répartitions en pourcentage dans les quatre provinces sont très voisines, ce qui justifie l'hypothèse de corrélations égales entre les termes d'erreur des modèles du recensement pour les différentes provinces. La similitude des répartitions permet également d'espérer une meilleure efficacité des estimateurs fondés sur le modèle, dans le cas du modèle conjoint, comparativement aux estimateurs qui ne tiennent pas compte des corrélations transversales entre les valeurs de la population pour les différentes provinces. Deux autres importants facteurs à considérer, mentionnés à la section 2.3, sont les suivants: le modèle devrait tenir compte d'effets possibles liés aux groupes de renouvellement, ainsi que des changements des variances des erreurs de l'enquête avec le temps. Pour obtenir des estimations

Tableau 1
Population active par industrie dans les provinces de l'atlantique, mars 1991

Taille de l'échantillon		Nouvelle-Écosse		Nouveau-Brunswick		Terre-Neuve		Île-du-Prince-Édouard	
Milliers		%		Milliers		%		Milliers	
409	100.0	7	1.7	311	100.0	233.0	100.0	61.0	100.0
Agriculture	7	1.7	7	2.3	0.5	0.2	0.2	6.0	9.8
Autres industries primaires	18	4.4	13	4.2	18.0	7.7	4.0	6.6	6.6
Secteur manufacturier	44	10.7	37	11.9	23.0	9.9	6.0	9.8	9.8
Construction	24	5.9	21	6.8	18.0	7.7	4.0	6.6	6.6
Transp. et communication	35	8.6	30	9.6	20.0	8.6	5.0	8.3	8.3
Commerce	81	19.8	61	19.6	41.0	17.6	10.0	16.4	16.4
Finances	20	4.9	12	3.9	6.0	2.6	0.5	0.8	0.8
Services	143	35.0	107	34.4	83.0	35.6	19.0	31.1	31.1
Administration publique	36	8.8	22	7.0	23.0	9.9	6.0	9.8	9.8
Non classé	1	0.2	1	0.3	0.5	0.2	0.5	0.8	0.8
Total	409	100.0	311	100.0	233.0	100.0	61.0	100.0	100.0

initiales des effets liés aux groupes de renouvellement, nous avons fait la moyenne des pseudo-erreurs de l'enquête, $e_{it}^{(j)} = (y_t^{(j)} - \hat{y}_t^{(j)}), j = 1, \dots, 6$, sur l'ensemble des mois de la période visée par l'échantillon. Nous avons ensuite divisé les moyennes par les estimations classiques des erreurs-types. (Les erreurs $e_{it}^{(j)}$ sont corrélées dans le temps, mais les corrélations sont faibles, car sauf pour les décalages 6, 12, etc., les données de n'importe quel panel urbain et à des secteurs de dénombrement différents dans les régions rurales. Voir la section 2.1.) Notons qu'en l'absence d'effets liés aux groupes de renouvellement, $E(e_{it}^{(j)}) = 0$ pour tous les j et t , peu importe le modèle postulé pour les valeurs de la population. Cette analyse préliminaire (indépendante du modèle) donne des résultats semblables aux résultats obtenus avec le modèle complet, présentés au tableau 2 de la section 3.3. Examinons maintenant les variances des erreurs de l'enquête. La figure 1 présente le graphique des effets saisonniers des estimateurs globaux de l'enquête dans les quatre provinces, ainsi que des effets saisonniers des erreurs-types de ces estimations (multipliés par 100). Comme précédemment, désignons par w_{ia} la taille relative de la population active dans la province a au temps t . L'estimateur global de l'enquête est défini par $y_t^* = \sum_{a=1}^4 w_{ia} y_{it}^{(a)}$ (équation 2.11). L'erreur-type de y_t^* est $(SD^*)_t = [\sum_{a=1}^4 w_{ia}^2 (SD_{it}^{(a)})^2]$. Les effets saisonniers ont été estimés par l'application du mode additif de la procédure X-1, afin qu'ils ne soient liés à aucun modèle particulier. Nous avons choisi le modèle additif puisque nous supposons une décomposition additive pour les estimateurs de l'enquête. (Comme le montre la figure 4, les effets saisonniers des estimateurs globaux de l'enquête produits par la procédure X-1 sont très voisins des effets saisonniers obtenus avec le modèle.) La figure 1 montre que les erreurs-types sont assujetties à des variations saisonnières, affichant un profil saisonnier qui est très voisin de celui des estimateurs de l'enquête et donc des valeurs correspondantes de la population.

i) Procéder à une détection des valeurs aberrantes de la série chronologique, comme il est proposé par exemple dans Chang, Tiao et Chen (1988).

ii) Modéliser les séries chronologiques des proportions $\{\pi_{ia} = y_{ia} / \sum_{a=1}^A y_{ia}, a = 1, \dots, (A - 1)\}$, si ces séries chronologiques affichent un profil plus lisse que les séries $\{y_{ia}\}$.

La détection des valeurs aberrantes est un important aspect de toute activité de modélisation, mais il reste à trouver la façon de modifier les estimations des valeurs de la population une fois que des observations (estimations de l'enquête) sont détectées à titre de valeurs aberrantes. Notons à ce propos que notre intérêt porte principalement sur les estimations courantes, c'est-à-dire les estimations les plus récentes disponibles. Dans Chang, Tiao et Chen (1988), la détection des valeurs aberrantes a pour objet de *retrancher* leur effet des observations, de façon à permettre de mieux comprendre la structure sous-jacente de la série et d'améliorer l'estimation des paramètres du modèle. Mais si la cause d'une valeur aberrante est un mouvement réel du niveau des valeurs de la population, ce mouvement ne doit pas être retranché, mais plutôt pris en compte dans les estimateurs fondés sur le modèle. Harrison et Stevens (1976) proposent de tenir compte de tels changements en modifiant la distribution antérieure des vecteurs d'états, p. ex. en accroissant les variances des erreurs des vecteurs d'états de façon à permettre des variations plus rapides des estimateurs des vecteurs d'états. Un exemple de cette façon de procéder est présenté dans Morris et Pfeffermann (1984). Notre méthode consistant à imposer la contrainte selon laquelle les estimateurs fondés sur le modèle doivent coïncider avec les estimateurs globaux de l'enquête offre un processus plus automatique, qui n'exige pas la modification de données antérieures.

La deuxième méthode suggérée pour aborder le problème de la robustesse est attrayante, car on peut s'attendre à ce que les variations brusques des valeurs de la population s'annulent dans les rapports π_{ia} . Le principal inconvénient du recours à cette méthode est que le modèle valable pour les rapports "vrais" π_{ia} est naturellement très différent du modèle représentant les valeurs de la population y_i tel qu'il est défini en (2.1) et que, notamment, il n'offre plus d'estimations de la tendance et des effets saisonniers, qui sont parmi les éléments les plus importants de notre approche, comme il est mentionné dans l'introduction. Il n'est pas évident, non plus, de trouver comment extraire les estimations des valeurs de la population Y_i du modèle valable pour les rapports π_{ia} sans faire des hypothèses additionnelles, comme par exemple notre hypothèse selon laquelle l'estimateur global de l'enquête est suffisamment voisin de la valeur correspondante de la population.

L'utilisation de contraintes de la forme (2.11) a été envisagée précédemment par Battese, Harter et Fuller (1988), ainsi que par Pfeffermann et Barnard (1991), pour l'analyse d'enquêtes transversales. Pfeffermann et Burck (1990) présentent des résultats empiriques illustrant la bonne performance des estimateurs modifiés au cours de périodes anormales. Voir aussi la section 3.

Notons que $\sum_{a=1}^A w_{ia} Y_{ia}$ et $\sum_{a=1}^A w_{ia} y_i$ sont respectivement l'estimateur fondé sur le modèle et l'estimateur direct de l'enquête de la valeur totale de la population pour le groupe de régions examiné. La condition 2.11 peut aussi être écrite sous la forme $\sum_{a=1}^A w_{ia} \tilde{e}_{ia} = 0$ où $\tilde{e}_{ia} = \sum_{j=1}^6 e_{ij}^{(a)} / 6$ est l'erreur de l'enquête moyenne pour la région a . Pfeffermann et Burck (1990) montrent comment modifier les équations du filtre de Kalman de façon qu'elles produisent l'estimateur des vecteurs d'états assujéti à cette contrainte et sa matrice V-C appropriée selon le modèle (sans la contrainte), pour chaque mois t .

La justification de cette modification est simple. On suppose que la taille totale de l'échantillon de l'ensemble des régions est suffisamment grande et, par conséquent, que les estimateurs globaux de l'enquête sont fiables. Cette hypothèse, en fait, dicte le niveau d'aggrégation requis; voir ci-dessous. En obligeant les estimateurs globaux fondés sur le modèle à coïncider avec les estimateurs globaux de l'enquête, l'analyste s'assure que toute variation réelle des valeurs de la population reflétée dans les estimateurs de l'enquête sera aussi reflétée dans les estimateurs fondés sur le modèle. Signaux qu'en l'absence d'une telle contrainte imposée aux estimateurs, s'il survenait par exemple des variations soudaines du niveau de la série, celles-ci ne se répercuteraient sur les estimateurs fondés sur le modèle que plusieurs mois plus tard, car ces estimateurs dépendent non seulement des données courantes, mais aussi des données passées. Par contre, si aucune variation prononcée ne survient, on peut s'attendre à ce que les estimateurs fondés sur le modèle se conforment approximativement aux contraintes, même si ces dernières ne sont pas imposées explicitement. Par conséquent, au cours de périodes normales, les estimateurs assujéti à la contrainte devraient être presque aussi performants que les estimateurs exempts de contrainte.

L'hypothèse selon laquelle la taille totale de l'échantillon de l'ensemble des régions est élevée et donc que l'estimateur global de l'enquête est suffisamment proche de la valeur correspondante de la population est critiquée. Elle garantit (probabilité élevée) que la modification interviendra uniquement s'il survient des changements réels dans les valeurs de la population, et non par suite de vastes erreurs d'échantillonnage. Il convient de signaler, comme l'a d'ailleurs mentionné l'un des arbitres, que dans l'application de la méthode aux provinces de l'Atlantique décrite à la section 3, l'estimateur global ne se fonde que sur quatre provinces, de sorte que son erreur-type est d'environ 50% des erreurs-types des estimateurs de l'enquête pour les provinces, selon les tailles des échantillons des provinces. (Les estimateurs de l'enquête pour les provinces sont conditionnellement indépendants, les conditions étant les valeurs de la population pour les provinces correspondantes.) Par conséquent, si les contraintes doivent être utilisées en pratique, l'aggrégation devrait être faite sur un plus grand nombre de provinces ou sur d'autres petites régions.

Les deux autres méthodes suivantes ont été suggérées pour aborder le problème de la robustesse:

chacune de ces réalisations, ce qui donne des estimations $\hat{q}_t(\hat{\lambda}^{(k)})$ avec matrices $V-C P_t(\hat{\lambda}^{(k)})$. Les matrices A_t et B_t sont ainsi estimées:

$$A_t = \frac{1}{K} \sum_{k=1}^K P_t(\hat{\lambda}^{(k)});$$
$$B_t = \frac{1}{K} \sum_{k=1}^K [\hat{q}_t(\hat{\lambda}^{(k)}) - \hat{q}_t(\hat{\lambda})][\hat{q}_t(\hat{\lambda}^{(k)}) - \hat{q}_t(\hat{\lambda})]'. \quad (2.9)$$

Ansley et Kohn (1986) proposent un estimateur pour B_t fondé sur une approximation de série de Taylor de premier ordre. L'utilisation de leur estimateur exige moins de calculs, mais la méthode proposée par Hamilton est un peu plus flexible sur le plan des hypothèses posées, et elle donne une meilleure idée de la mesure dans laquelle les résultats produits par le filtre de Kalman sont sensibles aux erreurs des estimateurs des paramètres.

2.6 Modélisation conjointe pour un groupe de petites régions

Le modèle examiné jusqu'ici concerne une seule région. Quand les tailles d'échantillon des diverses régions sont peu élevées, on peut souvent obtenir des estimateurs plus efficaces en modélisant en outre les relations transversales entre les valeurs de la population de plusieurs régions. De toute évidence, l'accroissement d'efficacité résultant d'une telle modélisation conjointe dépend de la taille des échantillons des petites régions et de la similitude, entre les régions, du comportement des valeurs de la population dans le temps.

Les erreurs de l'enquête étant indépendantes entre les régions, toute modélisation conjointe des estimateurs de l'enquête s'applique seulement au modèle du recensement. Pour modéliser les taux de chômage dans les quatre provinces de l'Atlantique, nous adoptons la démarche de Pfeffermann et Burck (1990) et permettons des corrélations contemporaines non nulles entre les termes d'erreurs correspondants des modèles du recensement valables dans ces provinces. Par conséquent, si $y_{it}^a = (e_{it}^{(a)}, \eta_{it}^{L(a)}, \eta_{it}^{R(b)})$ dénote le vecteur des termes d'erreur au temps t associé au modèle du recensement valable dans la région a , on suppose que $C_{a,b} = E(y_{it}^a y_{it}^b)$ est diagonale, mais comporte peut-être des covariances non nulles sur la diagonale principale. La conséquence réelle de cette hypothèse est que si, par exemple, il y a une augmentation sensible du niveau de la tendance dans une province, on peut s'attendre à ce que des augmentations semblables surviennent dans d'autres provinces.

Le modèle conjoint résultant, valable pour les quatre provinces (ou plus généralement pour un groupe de régions), peut encore être présenté sous forme d'un espace d'états; voir les équations (2.7) et (2.8) dans Pfeffermann et Burck (1990). L'ajustement de ce modèle pose toutefois un problème important, du fait que l'estimation conjointe de tous les paramètres inconnus exige trop de l'ordinateur

en temps de calcul et en espace de stockage. (Le programme informatique rédigé pour l'application de la méthode de notation utilise des dérivées du premier ordre numériques, de sorte que le calcul de chaque dérivée exige un balayage distinct de la totalité des données. Chaque balayage nécessite le calcul des équations du filtre de Kalman pour chaque mois inclus dans la période d'observation de l'échantillon.) Pour solutionner ce problème, nous avons d'abord ajusté les modèles définis par (2.5), (2.6) et (2.2) séparément pour chacune des provinces. Nous avons également supposé des corrélations égales entre les termes d'erreur correspondants des modèles du recensement des différentes provinces, c'est-à-dire

$$\phi_{a,b} = C_{a,b}^{-1/2} C_{a,b} C_{b,b}^{-1/2} = \phi \quad 1 \leq a, b \leq 4, \quad (2.10)$$

où $C_{a,a} = E(y_{it}^a y_{it}^a)$. Les quatre corrélations maximisant la vraisemblance du modèle conjoint ont été déterminées par une méthode de recherche par quadrillage, les autres paramètres du modèle étant maintenus fixes à leur valeur préalablement estimée.

L'hypothèse des corrélations égales réduit de beaucoup le nombre de paramètres inconnus. Elle se justifie aussi par le petit nombre de régions examinées dans cette étude, qui permet de supposer qu'aucune autre structure préexistante régissant ces corrélations ne pourrait être détectée de façon sûre. Plus concrètement, une simple ventilation de la population active par industrie (tableau 1 de la section 3) montre des fréquences relatives très semblables dans les quatre provinces, ce qui laisse croire à un degré élevé d'homogénéité de leurs économies.

2.7 Modifications pour se prémunir contre les défaillances du modèle

L'utilisation d'un modèle pour la production de statistiques officielles soulève la question de la protection contre d'éventuelles défaillances du modèle. Comme il est indiqué ci-après, il n'est pas possible de tester le modèle chaque fois que de nouvelles données deviennent disponibles. C'est pourquoi il faut doter le modèle d'un mécanisme intrinsèque visant à assurer la robustesse des estimateurs lorsque le modèle ne tient plus.

Pour la modélisation des séries sur la population active dans de petites régions, nous avons utilisé la modification proposée par Pfeffermann et Burck (1990). En vertu de cette modification, les estimations des vecteurs d'états mis à jour à n importe quel temps t sont soumises à la contrainte suivante:

$$\sum_{a=1}^A w_{it}^a x_{it}^a = \sum_{a=1}^A w_{it}^a y_{it}^a \quad t = 1, 2, \dots, \quad (2.11)$$

où x_{it}^a est l'estimateur fondé sur le modèle de la valeur de la population X_{it}^a dans la région a , $y_{it}^a = 1/6 \sum_{j=1}^6 y_{it}^{(j)a}$ est l'estimateur de l'enquête correspondant et $w_{it}^a = M_{it}^a/M_t$ est la taille relative de la population active dans cette région, c'est-à-dire que $M_t = \sum_{a=1}^A M_{it}^a$ et $\sum_{a=1}^A w_{it}^a = 1$.

La représentation du modèle sous forme d'espace d'états nous permet de mettre à jour, de lisser ou de prédire les vecteurs d'états et, par conséquent, la tendance, la composante saisonnière et la valeur de la population à n importe quel mois t , au moyen du filtre de Kalman. Désignons par \hat{q}_t le vecteur d'états correspondant au mois t . Le vecteur d'états comprend le niveau de la tendance, l'accroissement et les effets saisonniers, les biais liés aux groupes de renouvellement et les erreurs de l'enquête. Voir Pfeffermann (1991) pour plus de détails. Par "mise à jour", nous entendons l'estimation de \hat{q}_t au mois t , d'après toutes les données existant jusqu'au mois t inclusivement. Par "lissage", nous entendons l'estimation de \hat{q}_t d'après toutes les données disponibles pour l'ensemble des mois qui précèdent et qui suivent le mois t . Le lissage est nécessaire pour améliorer des estimations passées, par exemple lorsqu'on estime les effets saisonniers ou qu'on estime des variations des valeurs de la population ou de la tendance. Quant à la "prédiction" des vecteurs d'états relatifs à des mois postérieurs à la période d'observation de l'échantillon, elle est importante pour l'élaboration de politiques. Les prédictions portant sur la période d'observation de l'échantillon permettent d'évaluer la performance du modèle, p. ex. en comparant les estimations de panel découlant de la prédiction de vecteurs d'états avec les estimations réelles. Voir la section 3 pour plus de détails. La théorie des modèles d'espace d'états et du filtre de Kalman est élaborée dans plusieurs publications; voir Pfeffermann (1991) pour les équations de filtrage et de lissage, avec mention de références. Notons que les équations de filtrage et de lissage fournissent non seulement les trois ensembles d'estimateurs pour n importe quel mois donné t , mais aussi les matrices V-C des erreurs d'estimation correspondantes.

L'application réelle du filtre de Kalman exige l'estimation des paramètres inconnus du modèle et l'initialisation du filtre, c'est-à-dire l'estimation du vecteur d'états initial \hat{q}_0 et de la matrice V-C correspondante des erreurs d'estimation. Pour une petite région unique, les paramètres inconnus du modèle sont les quatre variances des termes d'erreur du modèle du recensement (2.1), ainsi que les huit coefficients d'autorégression et les six variances résiduelles des modèles des erreurs propres aux panels (2.2). (Les moyennes des groupes de renouvellement sont incluses dans les vecteurs d'états à titre de coefficients fixes, inva-riables dans le temps.) Afin de réduire le nombre de paramètres inconnus dans le modèle d'espace d'états combiné, nous supposons que $\sigma_j^2 = \sigma^2 \times \hat{\sigma}_j^2$, $j = 1, \dots, 6$, où les $\{\hat{\sigma}_j^2\}$ sont les variances résiduelles dans (2.2) et les $\hat{\sigma}_j^2$ sont les estimations des variances résiduelles obtenues par l'ajustement des équations d'autorégression aux pseudo-erreurs de l'enquête, $e_{i,j}^{(p)}$, définies à la section 2.3. Cette hypothèse réduit le nombre de paramètres inconnus de 18 à 13. (Les estimations $\hat{\sigma}_j^2$ sont très voisines pour $j = 4, 5, 6$ et ont été supposées égales.)

Pour estimer A_t et B_t , nous faisons porter la condition sur Y et nous suivons la méthode proposée par Hamilton (1986). Selon cette méthode, des réalisations $\hat{\lambda}^{(k)}$, $k = 1, \dots, K$ sont produites à partir de la distribution postérieure normale asymptotique de $\hat{\lambda}$, c.-à-d. à partir d'une distribution $N(\hat{\lambda}, \hat{\Lambda})$ où $\hat{\lambda}$ est l'estimateur du maximum de vraisemblance de $\hat{\lambda}$ et $\hat{\Lambda}$ est la matrice V-C asymptomatique de $\hat{\lambda}$. (Le filtre de Kalman est alors appliqué avec

$$\begin{aligned} Q_t &= E\{[\hat{q}_t(\hat{\lambda}) - \hat{q}_t][\hat{q}_t(\hat{\lambda}) - \hat{q}_t]'\} \\ &= E\{[\hat{q}_t(\hat{\lambda}) - \hat{q}_t][\hat{q}_t(\hat{\lambda}) - \hat{q}_t]'\} \\ &\quad + E\{[\hat{q}_t(\hat{\lambda}) - \hat{q}_t][\hat{q}_t(\hat{\lambda}) - \hat{q}_t]'\} \end{aligned} \quad (2.8)$$

Il s'agit de la somme de l'erreur qu'on obtiendrait si $\hat{\lambda}$ était connu et de l'erreur attribuable à l'estimation de $\hat{\lambda}$. Les deux termes du côté droit de (2.7) sont non corrélés. Un moyen simple de vérifier cette propriété consiste à observer que $\hat{q}_t(\hat{\lambda}) = E(\hat{q}_t | Y, \hat{\lambda})$, où Y représente l'ensemble des données disponibles. Puisque la condition porte sur Y et $\hat{\lambda}$, $E\{[\hat{q}_t(\hat{\lambda}) - \hat{q}_t] | Y, \hat{\lambda}\} = \hat{q}_t - \hat{q}_t(\hat{\lambda})$, et $\hat{\lambda}$, c'est-à-dire l'estimation du vecteur d'états initial du filtre, correspondante des erreurs d'estimation.

En notation mathématique, supposons que $\hat{q}_t(\hat{\lambda})$ soit l'estimateur de \hat{q}_t au mois t , d'après toutes les données disponibles jusqu'à un mois donné n , où $\hat{\lambda}$ représente les estimateurs des paramètres inconnus du modèle. L'erreur d'estimation peut être ainsi décomposée:

Une fois que les paramètres inconnus du modèle ont été estimés, les équations du filtre de Kalman peuvent être appliquées, les valeurs réelles des paramètres étant rem-placées par les estimations. Comme il a été indiqué à la section 2.4, le filtre de Kalman produit non seulement des estimations des vecteurs d'états, mais aussi des matrices V-C des erreurs d'estimation correspondantes. L'utilisation de ces matrices V-C peut toutefois poser un problème du fait que ces dernières ne tiennent pas compte de la variation additionnelle attribuable au fait que les paramètres sont estimés, ce qui peut engendrer une sous-estimation des variances vraies.

2.5 Ajustements pour tenir compte de l'utilisation de valeurs estimées des paramètres

Si l'on suppose que les termes d'erreur dans les modèles de recensement et des erreurs de l'enquête ont une distribution normale, les paramètres inconnus du modèle peuvent être estimés par une maximisation de la vraisemblance. Voir Pfeffermann et Burck (1991) pour une brève description de la façon d'utiliser l'algorithme de maximisation de la méthode de notation et pour des détails sur l'initialisation du filtre. Cet article inclut des références à des analyses plus approfondies.

chômage que pour celles de l'emploi, ce qui reflète la mobilité élevée de la population active en chômage. Les corrélations γ sont beaucoup plus faibles que les corrélations ρ , mais comme le signalé l'auteur, le calcul de ces corrélations est beaucoup moins fiable et leur comportement est quelque peu erratique, affichant parfois une tendance à la hausse. Nous avons calculé les corrélations sérielles d'après les modèles (2.2), en remplaçant les coefficients ϕ par leurs valeurs estimées et nous avons observé en général un ajustement étroit avec les corrélations ρ à tous les décalages (de 1 à 5). Les corrélations à des décalages plus élevés sont différentes des corrélations γ correspondantes, mais, fait intéressant, elles sont la plupart du temps plus élevées et diminuent toujours à mesure que j augmente.

Une autre question ayant trait au modèle (2.2) qu'ont soulevée les arbitres avait trait à la possibilité d'appliquer la transformation logarithmique aux données brutes de façon à stabiliser les variances des erreurs de l'enquête, plutôt que de modéliser les erreurs standardisées. Deux raisons principales incitent à ne pas recourir à la transformation logarithmique dans notre cas. En premier lieu, l'utilisation de cette transformation entraînerait une décomposition multiplicative des taux de chômage de la population, ce qui va à l'encontre de la pratique courante qui consiste à supposer une décomposition additive. À Statistique Canada, les taux de chômage des deux plus grandes provinces sur les quatre examinées dans notre étude sont désajustés sur la base d'une décomposition additive. Aux États-Unis, les modèles ajustés aux séries du chômage des États reposent eux aussi sur une décomposition additive. Voir Tiller (1992). La deuxième raison est que les changements des variances des erreurs de l'enquête peuvent résulter de modifications du plan d'échantillonnage et, en particulier, de changements des tailles des échantillons. De tels changements produisent des mouvements discrets des variances qui ne peuvent être traités efficacement par la transformation logarithmique. Comme l'a signalé par ailleurs l'un des arbitres, le fait de transformer les données à l'agrégation de produire une non-linéarité dans l'agrégation des estimations sur les divers panels et/ou petites régions.

Le modèle défini en (2.2) tient compte des deux facteurs à considérer énoncés plus haut. L'application réelle du modèle, toutefois, exige deux modifications:

1. Pour les trois premiers panels, il n'y a pas assez de données antérieures pour permettre l'ajustement d'un modèle AR(3). Par exemple, l'erreur de l'enquête pour correspond au panel qui fait partie de l'enquête pour la première fois. Pour surmonter ce problème, nous remplaçons les erreurs de l'enquête manquantes par les erreurs de l'enquête correspondant aux panels précédemment choisis dans les mêmes UPE ou strates. Par exemple, le modèle AR(2) ajusté à $e_{i(2)}^{(2)}$ est

$$e_{i(2)}^{(2)} = \phi_{21} e_{i(1)}^{(2)} + \phi_{22} e_{i(6)}^{(2)} + u_{i(2)}^{(2)} \quad (2.3)$$

ou

$$y_{i(j)}^{(j)} = L_i + S_i + \epsilon_i + e_{i(j)}^{(j)}, \quad j = 1, \dots, 6, \quad (2.5)$$

Il découle de (2.1) que les estimateurs des panels peuvent être modélisés par

$$L_i = L_{i-1} + R_{i-1} + \eta_{Li}; \quad R_i = R_{i-1} + \eta_{Ri};$$

$$S_{i+j} = \eta_{Si}, \quad \sum_{j=0}^f \quad (2.6)$$

2.4 Représentation sous forme d'espace d'états et

estimation du modèle représentant les estimateurs de l'enquête

où les coefficients γ sont estimés par les moindres carrés ordinaires. La notation $(SD)_i$ définit l'estimation brute, non lissée, de l'écart-type lié au plan de l'estimateur moyen de l'enquête, \bar{y}_i , au mois i et les $\{D_{it}\}$ sont des variables fictives tenant compte des effets saisonniers mensuels de telle sorte que $D_{it} = 1$ quand $t = 12k + i$, $k = 0, 1, \dots, 12$ et $D_{it} = 0$ dans les autres cas. Les écarts-types lissés des erreurs propres aux panels sont donnés par $\widehat{SD}(e_{i(j)}^{(j)}) = \sqrt{6(SD)_i}$. Ces dernières estimations sont utilisées comme substituts des écarts-types vrais, qui sont inconnus.

$$(SD)_i = \gamma(SD)_{i-1} + \gamma_0 i + \sum_{t=1}^{12} \gamma_t D_{it}, \quad (2.4)$$

2. Les écarts-types vrais des erreurs de l'enquête sont inconnus, tandis que les estimations de l'enquête des écarts-types sont elles-mêmes l'objet d'erreurs d'échantillonnage. Pour surmonter ce problème, nous utilisons des valeurs lissées des écarts-types estimés, obtenues par l'ajustement de la relation

avec $\{\epsilon_i\}$, $\{\eta_{Li}\}$, $\{\eta_{Ri}\}$ et $\{\eta_{Si}\}$ définis comme en (2.1). Les modèles distincts définis par (2.5), (2.6) et (2.2) peuvent être fondus en une représentation compacte d'espace d'états avec $\chi_i' = (y_{i(1)}^{(1)}, \dots, y_{i(6)}^{(6)})$ comme données d'entrée, semblable à la représentation décrite dans Pfeffermann (1991). Suivant cette représentation, les erreurs de l'enquête (et, dans la présente étude, les termes irréguliers du recensement également) sont inclus dans le vecteur d'états, de façon qu'il n'y ait pas de termes résiduels dans l'équation des observations définie par (2.5). Contrairement à la situation décrite dans Pfeffermann (1991), toutefois, la matrice de transition et la matrice de

Notons toutefois que dans le présent cas, la série $\{Y_t\}$ est elle-même non observable. Les séries $\{\eta_{Lt}\}$, $\{\eta_{Rt}\}$ et $\{\eta_{St}\}$ sont des perturbations de bruit blanc indépendantes de moyennes zéro et de variances σ_L^2 , σ_R^2 et $\sigma_S^2 \times g(t)$ respectivement. Par conséquent, la deuxième et la troisième équation de (2.1) définissent une approximation locale d'une tendance linéaire, tandis que la dernière équation décrit l'évolution des effets saisonniers de telle manière que la somme de n importe quel ensemble de 12 effets successifs fluctue autour de zéro. Notons que les variances des termes d'erreur η_{St} dépendent du temps. Les fonctions $g(t)$ sont définies à la fin de la section 3.1.

2.3 Le modèle des erreurs de l'enquête

Le modèle représentant les erreurs de l'enquête a été déterminé initialement par une analyse distincte des séries de pseudo-erreurs $e_{ij}^{(U)} = (y_{ij}^{(U)} - \bar{y}_i)$, $i = 1, \dots, N$, où $y_{ij}^{(U)}$ est l'estimateur de Y_i d'après le j -ième panel, $j = 1, \dots, 6$, (le panel observé pour le j -ième mois consécutif) et $\bar{y}_i = \sum_{j=1}^6 y_{ij}^{(U)}/6$ est l'estimateur moyen. Notons que sont les vraies erreurs de l'enquête. Ainsi, la caractéristique notable des contrastes $(y_{ij}^{(U)} - \bar{y}_i)$ est qu'ils sont fonction uniquement des erreurs de l'enquête, peu importe le modèle représentant les valeurs de la population.

Il y a deux facteurs principaux à considérer dans le choix d'un modèle pour les erreurs de l'enquête:

a) Le modèle devrait tenir compte d'éventuels biais liés aux groupes de renouvellement ou, plus généralement, permettre des moyennes différentes pour les erreurs propres à différents panels.

b) Le modèle devrait permettre que les variances des erreurs de l'enquête changent au fil du temps.

Des biais liés aux groupes de renouvellement peuvent survenir lorsque des répondants ne donnent pas les mêmes informations à des interviews différentes; ces effets peuvent dépendre de la durée de participation à l'enquête des répondants, ou de la méthode de collecte des données (p. ex. par téléphone ou par interviews à domicile). (Dans l'EPA canadienne, les membres du premier panel sont interviewés à domicile, tandis que ceux des autres panels sont interviewés par téléphone.) Une autre cause possible d'écarts entre les moyennes des erreurs propres aux panels réside dans les différences entre les profils de non-réponse d'un panel à l'autre. Voir Pfeffermann (1991) pour une analyse plus approfondie du problème et des références concernant les études antérieures effectuées dans ce domaine.

$$e_{ij}^{(U)} = \phi_{j1} e_{ij}^{(U-1)} + \phi_{j2} e_{ij}^{(U-2)} + \phi_{j3} e_{ij}^{(U-3)} + u_{ij}^{(U)}, j = 1, \dots, 6, \quad (2.2)$$

$SD(e_{ij}^{(U)})$, soit:

L'application de méthodes simples d'estimation de modèles et de diagnostic aux pseudo-erreurs de l'enquête suggère un modèle autorégressif (AR) de 3^e ordre pour les erreurs de l'enquête standardisées $e_{ij}^{(U)} = (y_{ij}^{(U)} - \bar{y}_i)/SD(e_{ij}^{(U)})$, soit:

où les $\beta_j = E(e_{ij}^{(U)})$ sont les biais liés aux groupes de renouvellement, les $SD(e_{ij}^{(U)})$ sont les écarts-types liés au plan et les $u_{ij}^{(U)}$ sont des termes de bruit blanc indépendants de moyenne zéro et de variances σ_j^2 . On suppose que $\sum_{j=1}^6 \beta_j = 0$, d'où il résulte que l'estimateur moyen de l'enquête, \bar{y}_i , est non biaisé. Voir Pfeffermann (1991) pour une analyse du besoin d'imposer une contrainte aux coefficients de biais. L'analyse subséquente de l'ajustement du modèle combiné défini par (2.1) et (2.2) (voir la section 2.4) permet de valider ce modèle, avec l'observation additionnelle selon laquelle les coefficients $(\phi_{j1}, \phi_{j2}, \phi_{j3})$ peuvent être supposés égaux pour $j = 4, 5, 6$. En outre, pour le premier panel, un modèle AR(1) donne déjà un bon ajustement, tandis que pour le deuxième et le troisième panel, un modèle AR(2) est approprié, bien que les coefficients soient différents. Ces relations sont valables pour chacune des quatre provinces de l'Atlantique.

L'un des arbitres ayant revu le présent article s'est demandé si le modèle AR(3) défini par (2.2) est suffisamment souple pour tenir compte des corrélations entre les estimations des panels à des décalages importants, qu'on croit être élevées en raison des "effets des UPE". Comme il a été mentionné à la section 2.1, les panels qui sortent de l'échantillon sont remplacés par des panels des mêmes UPE, et il faut habituellement plusieurs années avant qu'une UPE ne soit épuisée et remplacée par une UPE voisine. Lee (1990) présente deux ensembles de corrélations entre estimations de panels pour l'EPA canadienne. Le premier ensemble, dénoté par ρ_j , comprend les corrélations entre les estimations produites par le même panel, de sorte que j varie de 1 à 5. Le deuxième ensemble, dénoté par γ_j , comprend les corrélations entre les estimations produites par un panel et son prédécesseur, de sorte que j varie de 1 à 11. Les corrélations ρ sont généralement élevées, comme il faut s'y attendre, mais il convient de souligner qu'elles sont plus faibles pour les données du

2. UN MODÈLE D'ESPACE D'ÉTATS POUR LA SÉRIE CHRONOLOGIQUE DES DONNÉES SUR LE CHÔMAGE AU CANADA

2.1 L'enquête sur la population active du Canada

Les données canadiennes sur le chômage sont recueillies dans le cadre de l'enquête sur la population active (EPA) réalisée par Statistique Canada. L'EPA canadienne est une enquête par panel mensuelle avec renouvellement, dans laquelle chaque nouveau panel de ménages observé demeure dans l'échantillon pendant six mois consécutifs avant d'être remplacé par un autre panel de la même unité primaire d'échantillonnage (UPÉ) ou de la même strate. Les UPÉ sont délimitées géographiquement (pâtés de maisons ou centres urbains dans les régions urbaines, et groupes de secteurs de dénombrement dans les régions rurales). Les strates sont des groupes homogènes d'UPÉ délimités géographiquement (p. ex. secteurs de recensement, subdivisions de recensement et secteurs de dénombrement). Dans les régions rurales, les UPÉ sont représentées dans un seul panel. Dans les régions urbaines, chaque UPÉ est représentée dans un seul panel. Par conséquent, les estimateurs de panel distincts peuvent être présumés indépendants, propriété qui a été validée et utilisée dans d'autres études; voir par exemple Lee (1990). Pour une description récente du plan de l'EPA et de la construction des estimateurs directs de l'enquête, le lecteur est invité à consulter Singh et coll. (1990).

2.2 Le modèle du recensement

Dans les paragraphes qui suivent, nous examinons le cas d'une petite région unique, tandis que dans la section 2.4, nous aborderons la modélisation conjointe des estimations des panels pour un groupe de petites régions. Le modèle postulé pour les valeurs de la population est le modèle structural de base (MSB), représenté par l'ensemble suivant d'équations:

$$\begin{aligned} Y_t &= L_t + S_t + \epsilon_t; \quad L_t = L_{t-1} + R_{t-1} + \eta_{Lt}; \\ R_t &= R_{t-1} + \eta_{Rt}; \quad \sum_{i=1}^J S_{t+i} = \eta_{St}. \end{aligned} \tag{2.1}$$

Dans 2.1, Y_t est la valeur de la population (taux de chômage, "vrai") au temps t , L_t est le niveau de la tendance, R_t est l'accroissement, S_t est l'effet saisonnier et ϵ_t est le terme irrégulier qu'on suppose être un bruit blanc avec moyenne zéro et variance σ_ϵ^2 . Ainsi, la première équation de (2.1) suppose la décomposition classique d'une série chronologique en trois composantes: tendance, variations saisonnières et terme irrégulier. Une telle décomposition est inhérente aux méthodes courantes de désaisonnalisation des données; voir par exemple Dagum (1980).

i) elle facilite la détermination du modèle de série chronologique reflétant les erreurs de l'enquête, en permettant l'analyse des contrastes entre les estimateurs des différents panels, et ii) elle donne des estimateurs plus efficaces pour les paramètres du modèle, et donc de meilleurs prédicteurs des composantes non observables du modèle.

3) Le modèle tient compte des changements des variances des erreurs de l'enquête au fil du temps, ainsi que des effets possibles des groupes de renouvellement.

4) Le modèle peut être appliqué simultanément aux estimateurs des panels de petites régions distinctes. Dans ce cas, le modèle du recensement est étendu de manière à tenir compte des corrélations croisées entre les composantes non observables des valeurs de la population qui existent dans ces régions.

5) Une modification visant à assurer la robustesse des estimateurs des petites régions à l'égard d'éventuelles défaillances du modèle est incorporée aux équations du modèle. La modification consiste à contraindre les estimateurs fondés sur le modèle d'agrégats de valeurs de la population pour un groupe de petites régions dont la taille totale de l'échantillon est suffisamment grande à coïncider avec les estimateurs de l'enquête pour les agrégats correspondants. Ainsi, les variations soudaines du niveau de la série sont répercutées sans retard sur les estimateurs fondés sur le modèle.

Le modèle, ainsi que les modifications apportées pour en assurer la robustesse, sont décrits de façon plus détaillée à la section 2. Les résultats empiriques de l'application du modèle aux quatre provinces de l'Atlantique du Canada sont présentés à la section 3. La section 4 présente un bref sommaire ainsi que des suggestions quant à la direction que pourraient emprunter les recherches futures.

Avant de conclure la présente section, mentionnons qu'aux États-Unis, les estimations du chômage sont produites pour la plupart des États à l'aide de modèles de série chronologique ayant une structure semblable à celle du modèle de notre étude. Voir Tiller (1992) pour plus de détails. Une différence importante entre les deux, toutefois, est qu'aux États-Unis, le modèle postulé pour les valeurs de la population inclut également des variables explicatives, de sorte que la tendance et la composante saisonnière ne reflètent que la tendance et les variations saisonnières non prises en compte par les variables explicatives. Les modèles ajustés aux erreurs de l'enquête sont, comme dans notre cas, du type ARMMI, et ils tiennent compte eux aussi des changements des variances des erreurs de l'enquête. Les modèles sont par ailleurs différents en raison des modes très distincts de renouvellement de l'échantillon utilisés dans les deux pays. Une autre différence notable entre les deux modèles vient du fait qu'aux États-Unis, les modèles sont ajustés à chaque État séparément et que les données d'entrée comprennent seulement les estimations moyennes de l'enquête, c'est-à-dire une seule observation chaque mois. Les modèles, par conséquent, ne tiennent pas compte des biais liés aux groupes de renouvellement.

Modélisation conjointe robuste de séries de données sur l'activité pour de petites régions

D. PFEFFERMAN et S.R. BLEUER¹

RÉSUMÉ

Dans cet article, nous présentons les résultats de l'application d'un modèle d'espace d'états aux taux de chômage canadiens. Le modèle suppose une décomposition additive des valeurs de la population en une tendance, une composante saisonnière et une composante irrégulière, ainsi que des relations autorégressives distinctes pour les six séries d'erreurs de l'enquête correspondant aux six estimateurs de panel mensuels. Le modèle tient compte des effets des groupes de renouvellement et permet que changent, dans le temps, les variances liées au plan qui affectent les erreurs de l'enquête. L'ajustement du modèle est effectué au niveau de petites régions, mais il tient compte de corrélations entre les séries des composantes pour différentes régions. On obtient la robustesse des estimateurs produits par le modèle en imposant, à titre de contrainte, que les estimateurs globaux mensuels fondés sur le modèle visant un groupe de petites régions pour lequel la taille d'échantillon totale est suffisamment grande coïncident avec les estimateurs directs correspondants de l'enquête. La performance du modèle, dans le cas d'une application aux provinces de l'Atlantique, est évaluée au moyen de diverses statistiques de diagnostic et de graphiques de résidus, ainsi que par des comparaisons avec des estimateurs actuellement en usage.

MOTS CLÉS: Variance liée au plan; filtre de Kalman; enquête par panel; biais lié au renouvellement; modèle d'espace d'états.

Un modèle de série chronologique appliqué à des données d'enquête est la combinaison de deux modèles distincts: le "modèle du recensement", qui décrit l'évolution dans le temps des valeurs de la population finie, et le modèle des erreurs de l'enquête, qui décrit les relations qui existent dans la série chronologique des erreurs des estimateurs de l'enquête. Il y a au moins quatre raisons principales qui incitent à modéliser les estimateurs bruts de l'enquête:

a) Les estimateurs des valeurs de la population qui résultent du processus de modélisation (estimateurs fondés sur le modèle) ont en général des variances moindres que les estimateurs de l'enquête, notamment dans les petites régions, où les tailles des échantillons sont peu élevées.

b) Le modèle que nous employons permet d'obtenir des estimateurs pour les effets saisonniers, ainsi que pour les variances de ces estimateurs, comme sous-produit du processus d'estimation.

c) Le modèle permet d'établir des prévisions des valeurs saisonnières à l'égard de périodes postérieures à la période d'observation de l'échantillon pour laquelle les estimateurs directs de l'enquête sont disponibles. De

telles prévisions sont importantes pour l'évaluation de la performance du modèle, de même que pour l'élaboration de politiques.

d) Le modèle peut servir à détecter des points tournants dans le niveau des séries chronologiques et à en évaluer les conséquences. (Les travaux réalisés dans ce domaine seront exposés dans un article distinct.)

La méthodologie décrite dans le présent article intègre les méthodologies présentées dans Pfeffermann et Burck (1990) et Pfeffermann (1991), avec certaines modifications et extensions. Les principales caractéristiques du modèle sont les suivantes:

1) Le modèle décompose les valeurs de la population en trois composantes non observables: tendance, saisonnalité et termes irréguliers. Des prédicteurs lissés de ces composantes (et donc des valeurs de la population) fondés sur l'ensemble des données disponibles, ainsi que les erreurs-types des erreurs de prédiction, sont obtenus directement par l'application du filtre de Kalman. Les erreurs types sont modifiées pour tenir compte de la variation additionnelle causée par l'utilisation de valeurs estimées de paramètres.

2) Le modèle utilise comme données d'entrée les différents estimateurs de panel mensuels. L'utilisation des estimateurs propres à chaque panel offre deux avantages importants par rapport à l'emploi des estimateurs moyens:

¹ D. Pfeffermann, Department of Statistics, Hebrew University, Jerusalem 91905; S.R. Bleuer, Division des méthodes d'enquêtes sociales, Statistique Canada, Ottawa, Ontario, K1A 0T6.

que de la présentation. Nous désirons aussi remercier Martha Fair et Pierre Lalonde qui nous ont permis d'obtenir les données sur les mineurs ontariens ainsi que les données sur la fréquence des résultats pour les concordances vraies.

BIBLIOGRAPHIE

- BARTLETT, S., KREWSKI, D., WANG, Y., et ZIELINSKI, J.M. (1992). Évaluation des taux d'erreur dans de grandes études par couplage d'enregistrements informatisés. *Techniques d'enquête*, 19, 3-13.
- BELIN, T.R. (1990). A proposed improvement in computer matching techniques. Dans *Statistics of Income and Related Administrative Record Research: 1988-1989*, U.S. Internal Revenue Service, 167-172.
- BELIN, T.R., et RUBIN, D.B. (1991). Recent developments in calibrating error rates for computer matching. *Proceedings of the Annual Research Conference, U.S. Bureau of the Census*, 657-668.
- FAIR, M.E., et LALONDE, P. (1987). Identificateurs manquants et justesse de l'observation suivie. *Recueil: Les utilisations statistiques des données administratives, Statistique Canada*, 111-125.
- FAIR, M.E., NEWCOMBE, H.B., et LALONDE, P. (1988). Improved mortality searches for Ontario miners using social insurance index identifiers. Rapport de recherche, Commission de contrôle de l'énergie atomique.
- FELLEGI, I.P., et SUNTER, A.B. (1969). A theory for record linkage. *Journal of the American Statistical Association*, 64, 1183-1210.
- HABERMAN, S.J. (1976). Iterative scaling procedures for log-linear models for frequency tables derived by indirect observation. *Proceedings of the Statistical Computing Section, American Statistical Association*, 45-50.
- HABERMAN, S.J. (1979). *Analysis of Qualitative Data*. London: Academic Press.
- IMSL (1987). *Math/Library FORTRAN subroutines for mathematical applications*. Houston: IMSL Inc.
- MORE, J., GARBOW, B., et HILLSTROM, K. (1980). User guide for MINPACK-1. Argonne National Labs Report ANL-80-74.
- NEWCOMBE, H.B., KENNEDY, J.M., AXFORD, S.J., et JAMES, A.P. (1959). Automatic linkage of vital records. *Science*, 130, 954-959.
- NEWCOMBE, H.B. (1988). *Handbook of Record Linkage: Methods for Health and Statistical Studies, Administration, and Business*. Oxford: Oxford University Press.
- THIBAUDEAU, Y. (1989). Fitting log-linear models in computer matching. *Proceedings of the Statistical Computing Section, American Statistical Association*, 283-288.
- WINKLER, W.E. (1989). Near automatic weight computation in the Fellegi-Sunter model of record linkage. *Proceedings of the Annual Research Conference, U.S. Bureau of the Census*, 145-155.

cette dernière méthode.

La troisième méthode est basée sur un paramétrage de dépendances entre les résultats de comparaisons, pour différents champs de données, à l'aide d'effets log-linéaires. Avec ce paramétrage, on peut obtenir des estimations des probabilités d'accord qui ne sont pas fondées sur l'hypothèse d'indépendance en utilisant la méthode de pondération itérative pour estimer les paramètres d'un modèle à variable latente. Pour les ensembles de données synthétiques, avec absence d'indépendance, les estimations des taux d'erreur de classification fondées sur un modèle calculé par pondération itérative comprenaient des biais de beaucoup inférieurs à ceux qui s'appliquaient aux estimations basées sur l'hypothèse d'indépendance. Bien que l'ajustement obtenu à l'aide du modèle à variable latente pour la plupart des ensembles de données synthétiques ait été meilleur que celui obtenu pour un modèle fondé sur l'hypothèse d'indépendance, on relevait parfois un manque d'ajustement important avec le premier modèle. Quand les données synthétiques ont été modifiées afin d'améliorer l'ajustement obtenu avec le modèle à variable latente, rien ne montrait qu'il y avait un biais dans les estimations du taux d'erreur de classification fondées sur un modèle. Pour les données réelles, il y avait des écarts importants par rapport à l'hypothèse d'indépendance tant pour les concordances vraies que pour les non-concordances vraies. Les estimations vraies que pour les non-concordances vraies.

Les résultats présentés ici montrent que l'on peut améliorer les propriétés des estimations du taux d'erreur de classification fondées sur un modèle quand on utilise un estimateur approprié des probabilités d'accord. Les modèles à variable latente ainsi que la pondération itérative fournissent une méthode pour incorporer des dépendances entre des résultats de comparaisons pour différents champs de données pendant l'estimation des probabilités d'accord.

REMERCIEMENTS

Les auteurs désirent remercier William Winkler qui a fourni le code machine sur lequel est basé le programme d'estimation par pondération itérative que nous avons utilisé pour obtenir nos résultats, ainsi que Fritz Scheuren et trois arbitres anonymes pour leurs commentaires, sur une version antérieure de cet article, qui ont mené à des améliorations considérables tant pour ce qui est du contenu

Tableau 7
Taux d'erreur de classification, non-concordances vraies, données réelles

Taux estimé ($\times 12,84$)	Méthode des moments	Méthode itérative $\mu^0 = 0,0000625$	Méthode itérative $\mu^0 = 0,00025$	Méthode itérative $\mu^0 = 0,001$	Pondération itérative
Taux réel ($\times 12,84$)					
0.02	0.0368	1.311	0.1859	0.186	0.0339
0.04	0.0796	1.314	0.1888	0.193	0.0649
0.06	0.1224	1.317	0.1917	0.1967	0.0684
0.08	0.1573	1.323	0.1990	0.1994	0.1106
0.10	0.1863	1.333	0.60	0.4066	0.1282

Tableau 8
Taux d'erreur de classification, concordances vraies, données réelles

Taux estimé	Méthode des moments	Méthode itérative $\mu^0 = 0,0000625$	Méthode itérative $\mu^0 = 0,00025$	Méthode itérative $\mu^0 = 0,001$	Pondération itérative
Taux réel					
0.02	0.0166	0.0141	0.0193	0.0225	0.0105
0.04	0.0318	0.0264	0.029	0.0278	0.0263
0.06	0.0598	0.0383	0.0472	0.0326	0.0529
0.08	0.0782	0.0416	0.1372	0.0488	0.0784
0.10	0.0966	0.045	0.1393	0.1371	0.0958

les non-concordances vraies.
estimations obtenues avec la méthode des moments pour vraies. Toutefois, elles sont beaucoup plus précises que les pondération itérative sont un peu moins précises que celles basées sur la méthode des moments pour les concordances vraies. Toutfois, elles sont beaucoup plus précises que les estimations obtenues avec la méthode des moments pour les non-concordances vraies. Bien que le modèle $U(1)$, $U(2)U(4)$, $U(3)U(4)$ ne décrive pas de façon adéquate les dépendances parmi les non-concordances vraies, l'algorithme de pondération itérative a permis d'obtenir un bon ajustement à l'aide d'une estimation de la proportion d'enregistrements apparités (0.0747) qui diffère un peu de la valeur vraie (0.0722). On peut aussi obtenir un ajustement semblable à l'aide du modèle $G(1)G(2)$, $G(1)G(3)$, $G(4)$ ainsi qu'une estimation de 0.077 pour la proportion des appartements. Les estimations du taux d'erreur basées sur le modèle $G(1)G(2)$, $G(1)G(3)$, $G(4)$ ne sont pas meilleures que les estimations obtenues à l'aide de la méthode des moments.

7. CONCLUSIONS

Dans cet article, on a traité de la question de l'estimation des taux d'erreur de classification pour le couplage d'enregistrements. Le modèle de Fellegi-Sunter permet de calculer des estimations des taux d'erreur de classification

à l'aide d'estimations des probabilités d'accord. Ces estimations fondées sur un modèle ont généralement de mauvaises propriétés en pratique. Il a été démontré que leurs propriétés peuvent être améliorées en estimant avec soin les probabilités d'accord. Trois méthodes d'estimation ont été évaluées à l'aide de données synthétiques ainsi que de renseignements provenant d'une application réelle.
Pour deux des trois méthodes, on a utilisé l'hypothèse que les résultats des comparaisons pour différents champs de données sont indépendants. Cette hypothèse n'était pas valable soit pour les données synthétiques, soit pour les données réelles. Les données synthétiques incluaient de fortes dépendances pour les concordances vraies et des dépendances mineures pour les non-concordances vraies. Les dépendances pour les données réelles étaient particulièrement importantes dans le cas des non-concordances vraies. Les estimations du taux d'erreur de classification obtenues à l'aide de la méthode des moments, qui est fondée sur l'hypothèse d'indépendance, comportaient des biais considérables pour les données synthétiques et étaient relativement imprécises pour les données réelles. L'importance du biais dans les estimations des taux d'erreur de classification obtenues à l'aide de la méthode itérative dépendait de la définition d'un ensemble initial de concordances. Bien que certaines définitions de l'ensemble initial de concordances aient mené à des biais relativement faibles, d'autres ont produit des estimations avec des biais

particuliers pendant la période allant de 1964 à 1977 inclusivement. Le fichier des mineurs n'incluait que les enregistrés avec un numéro d'assurance sociale valable. Le deuxième fichier renfermait des enregistrés qui avaient été retenus après une comparaison initiale visant à éliminer les enregistrés qui n'avaient aucune similitude avec un quelconque des enregistrés dans les fichiers des mineurs. Le statut vital de chaque mineur, à la fin de 1977, avait été classé comme "décès confirmé", "survie confirmée" ou "non retrouvé lors du suivi", basé sur un couplage antérieur, combiné avec des procédures de suivi complètes, y compris un examen manuel. Les enregistrés dans le fichier des mineurs, pour les personnes dont le statut vital est "décès confirmé", incluaient le numéro d'enregistrement du décès dans la BCDM. On peut trouver plus de renseignements sur la construction des fichiers et sur les procédures utilisées pour déterminer l'état véritable du couplage dans Fair et Lalonde.

naissance et le mois de naissance, ont été choisis comme champs d'appariement pour la comparaison. Les enregistrements dans le fichier des mineurs pour lesquels le statut vital était "non retrouvé lors du suivi" ont été éliminés. Après que les enregistrements pour lesquels des valeurs manquaient soit dans au moins un champ d'appariement, soit dans le champ de l'année de naissance, aient aussi été supprimés, le fichier A (basé sur le fichier des mineurs)

Les fréquences des résultats pour les concordances vraies et pour les non-concordances vraies sont présentées au tableau 6. Tous les modèles log-linéaires correspondant à un modèle à variable latente non saturé (c'est-à-dire, tous les modèles comprenant moins de huit paramètres) sont rejetés par les données sur les fréquences pour les non-concordances vraies à un niveau de signification très faible. Parmi les modèles comprenant moins de huit paramètres, le modèle $U(1)$, $U(2)U(4)$, $U(3)U(4)$ correspond à la statistique la plus faible du test du rapport des vraisemblances pour le manque d'ajustement, soit 35,29. Le modèle $M(1)$, $M(2)M(4)$, $M(3)M(4)$ fournit un ajustement adéquat avec les données pour la concordance vraie (la statistique du test du rapport des vraisemblances est 10,29).

à l'aide de la méthode des moments, de la méthode itérative et de la pondération itérative, en utilisant le modèle à variable latente $G(1), G(2)G(4), G(3)G(4)$. La statistique du test du rapport des vraisemblances pour le modèle fondé sur l'hypothèse d'indépendance correspondant à l'estimateur de la méthode des moments est 108 (six degrés de liberté). Le modèle fondé sur l'hypothèse d'indépendance est rejeté par les données à un niveau de signification très

Tableau 6

Résultat selon l'indicateur: 0 = Désaccord, 1 = Accord		Fréquence	
Code NYSIS	Code du nom de jeune fille de la mère	Mois de naissance	Non- concor- dances vraies
Premier prénom		de naissance	Concor- dances vraies
0	0	0	4
0	0	0	3
0	0	1	11
0	0	1	128
0	1	0	3
0	1	0	7
0	1	1	27
0	1	1	242
1	0	0	9
1	0	0	10
1	0	0	52
1	0	1	392
1	1	0	27
1	1	0	32
1	1	0	115
1	1	1	1,001
Total			2,063
			26,500

Les estimations, fondées sur un modèle, du taux d'erreur de classification obtenues à l'aide de la méthode itérative sont très imprécises, particulièrement pour les non-concordances vraies, quelle que soit la valeur de p_0^0 . Les estimations du taux d'erreur obtenues à l'aide de la

Tableau 3

Taux d'erreur de classification, concordances vraies, données synthétiques (erreur-typés Monte Carlo entre parenthèses)				
Taux réel				
Taux estimé	Méthode des moments	Méthode itérative $\mu^0 = 0.000625$	Méthode itérative $\mu^0 = 0.00025$	Méthode itérative $\mu^0 = 0.001$
0.02	0.0580 (0.0013)	0.1179 (0.0041)	0.0507 (0.0014)	0.0149 (0.0008)
0.04	0.0773 (0.0014)	0.1362 (0.004)	0.0735 (0.0012)	0.0359 (0.0018)
0.06	0.0966 (0.0014)	0.1542 (0.0038)	0.0954 (0.0012)	0.0660 (0.0014)
0.08	0.1159 (0.0014)	0.1722 (0.0036)	0.1165 (0.0012)	0.0866 (0.0017)
0.10	0.1348 (0.0014)	0.1904 (0.0035)	0.1319 (0.0014)	0.1025 (0.002)
				0.1043 (0.0019)
				0.0841 (0.0018)
				0.0646 (0.0016)
				0.0455 (0.0012)
				0.025 (0.002)

Tableau 4

Taux d'erreur de classification, non-concordances vraies, données synthétiques modifiées (erreur-typés Monte Carlo entre parenthèses)				
Taux réel ($\times 99$)				
Taux estimé ($\times 99$)	Méthode des moments	Pondération itérative		
0.02	0.0189 (0.0008)	0.0194 (0.001)		
0.04	0.0385 (0.0011)	0.0396 (0.0016)		
0.06	0.0577 (0.0013)	0.0589 (0.0019)		
0.08	0.0767 (0.0016)	0.0785 (0.002)		
0.10	0.0957 (0.002)	0.0978 (0.0021)		

Tableau 5

Taux d'erreur de classification, concordances vraies, données synthétiques modifiées (erreurs-typés Monte Carlo entre parenthèses)				
Taux réel				
Taux estimé	Méthode des moments	Pondération itérative		
0.02	0.0553 (0.0014)	0.0208 (0.0011)		
0.04	0.0747 (0.0014)	0.0415 (0.0016)		
0.06	0.094 (0.0014)	0.0608 (0.0018)		
0.08	0.1134 (0.0014)	0.0805 (0.002)		
0.10	0.1325 (0.0015)	0.1007 (0.002)		

6. COMPARAISON DES MÉTHODES D'ESTIMATION - DONNÉES RÉELLES

Les résultats des comparaisons des trois méthodes d'estimation effectuées à l'aide de données provenant d'une application de couplage d'enregistrements sont présentées dans cette section. Deux fichiers de données utilisés dans un travail empirique présenté par Fair et Lalonde (1987) ont été employés. Le premier fichier renfermait des renseignements sur les mineurs ontariens obtenus de la Commission des accidents du travail. Le deuxième fichier comprenait des renseignements tirés de la Base canadienne de données sur la mortalité (BCDM) pour les décès de

Les renseignements présentés dans les tableaux 4 et 5 sont basés sur une série d'ensembles de données synthétiques produits à l'aide d'une version modifiée du tableau 1. Les valeurs probables des chiffres dans les cases du tableau 1 selon le modèle $M(1)M(2), M(3), M(4)$ ont été utilisées pour produire les données. Les biais dans les estimations du taux d'erreur de classification fondées sur un modèle obtenues à l'aide de la méthode des moments sont fortement réduits quand on utilise le modèle à variable latente $G(1)G(2), G(3), G(4)$ estimé à l'aide de la pondération itérative, particulièrement dans le cas des concordances vraies.

Tableau 2

Taux d'erreur de classification, non-concordances vraies, données synthétiques (erreurs-types Monte Carlo entre parenthèses)

Taux réel (× 99)

Taux estimé (× 99)	Méthode des moments	Méthode itérative $\mu^0 = 0.000625$	Méthode itérative $\mu^0 = 0.00025$	Méthode itérative $\mu^0 = 0.001$	Pondération itérative
0.02	0.0188 (0.0008)	0.0208 (0.0008)	0.0208 (0.001)	0.0207 (0.001)	0.0195 (0.001)
0.04	0.0381 (0.001)	0.0408 (0.0013)	0.0407 (0.0016)	0.0405 (0.0016)	0.0397 (0.0016)
0.06	0.057 (0.0012)	0.0626 (0.0015)	0.0615 (0.0018)	0.0602 (0.0019)	0.059 (0.0018)
0.08	0.076 (0.0015)	0.0855 (0.0017)	0.0838 (0.0019)	0.0804 (0.0022)	0.0785 (0.0019)
0.10	0.095 (0.0019)	0.1086 (0.0021)	0.1061 (0.0022)	0.1007 (0.0026)	0.0978 (0.0021)

Les propriétés de la méthode itérative dépendent des définitions des ensembles initiaux de concordances et de non-concordances, M^0 et U^0 . Il faut se rappeler que, compte tenu de probabilités initiales, les paires d'enregistrements sont classées selon la règle suivante:

$$\begin{aligned} & j \in M^0 \quad \text{si} \quad \omega^j > \tau_1^0, \\ & j \in U^0 \quad \text{si} \quad \omega^j < \tau_2^0. \end{aligned}$$

Quand la méthode itérative a été appliquée pour les simulations mentionnées ici, on a donné à τ_2^0 la valeur de τ_1^0 . Pour chaque essai de Monte Carlo, τ_1^0 a été fixé de façon à ce que

$$P(j \in U \mid \omega^j > \tau_1^0) + \gamma \cdot P(j \in U \mid \omega^j = \tau_1^0) = \mu^0,$$

pour un $\gamma \in [0, 1)$, où les probabilités estimées sont basées sur les estimations itératives initiales de \bar{y} . Les paires d'enregistrements avec poids τ_1^0 ont été classifiées dans M^0 avec probabilité γ . C'est-à-dire que l'ensemble initial de concordances utilisé par la méthode itérative était défini de façon à ce que le taux de fausses concordances estimé correspondant soit μ^0 . Les valeurs initiales pour m_k , $k = 1, 2, \dots, 4$, ont été fixées à 0.9.

Le chiffre zéro dans le tableau 1 (accord pour le prénom, désaccord pour tous les autres identificateurs) était traité comme un zéro structurel pendant la production des données. Parmi les modèles log-linéaires pour lesquels on n'utilisait pas plus de six paramètres, c'est le modèle $M(1)M(2), M(3), M(4)$ qui donne le meilleur ajustement avec les données du tableau 1. Ce modèle, dans lequel on utilise la dépendance pour les résultats de comparaisons pour le prénom et l'initiale d'un deuxième prénom, ne donne pas un très bon ajustement. La statistique du test du rapport des vraisemblances pour le manque d'ajustement

Les moyennes des estimations du taux d'erreur de classification obtenues à l'aide des ensembles de données synthétiques ainsi que les erreurs-types de Monte Carlo correspondantes sont présentées au tableau 2 pour les non-concordances vraies et au tableau 3 pour les concordances vraies. Après multiplication par 99, les taux d'erreur pour les non-concordances vraies mal classées divisé par le nombre de concordances vraies. Les résultats sont présentés pour la méthode des moments et pour la pondération itérative, ainsi que pour la méthode itérative avec $\mu^0 = 0.000625$, 0.00025 et 0.001. Les biais dans les taux d'erreur estimés pour les non-concordances vraies sont généralement faibles. La méthode itérative avec $\mu^0 = 0.001$ fournit les meilleures estimations, vient ensuite la pondération itérative. Pour les concordances vraies, le rendement de la méthode itérative dépend beaucoup du choix de μ^0 . Bien que la méthode itérative donne de bons résultats pour $\mu^0 = 0.001$, les biais pour $\mu^0 = 0.000625$ et $\mu^0 = 0.00025$ sont considérables. Les estimations du taux d'erreur de classification pour les concordances vraies obtenues à l'aide de la méthode des moments comprennent aussi des biais importants. Les biais dans les estimations obtenues par pondération itérative sont relativement faibles.

limité aux 50 prénoms francophones les plus courants et aux 50 prénoms non francophones les plus courants. On n'a pas tenu compte des variantes orthographiques lors du choix des noms. On a toutefois pris en considération toutes les initiales d'un deuxième prénom et de toutes les années de naissance qui figuraient dans le fichier pour 1988. La probabilité que des enregistrements avec des prénoms anglophones se voient attribuer un nom de famille anglo-phones (ce qui reflète la répartition des noms dans la population canadienne), à part ces restrictions, les identificateurs étaient choisis indépendamment.

Le point de départ pour le fichier B était une copie exacte du fichier A. Pour chaque enregistrement du fichier B, il y avait une concordance vraie avec exactement un enregistrement du fichier A. Pour introduire une absence d'indépendance parmi les concordances vraies, on a tiré un vecteur de résultats de la distribution de fréquences du tableau 1 pour chaque enregistrement du fichier B. Les identificateurs correspondant à des zéros dans le vecteur de résultats ont été choisis à nouveau. Par conséquent, l'ensemble de vecteurs de résultats pour les concordances vraies était un échantillon de la distribution du tableau 1. Les ensembles de données synthétiques comprenaient aussi de légers écarts par rapport à l'hypothèse d'indépendance pour les non-concordances vraies puisque la sélection des prénoms et des noms de famille n'était pas complètement indépendante.

Chaque ensemble de résultats de simulations mentionné plus loin est basé sur 50 essais de Monte Carlo. Chaque essai comportait la production de fichiers A et B comprenant 500 personnes, l'estimation de \bar{m} et de \bar{y} , la détermination de seuils correspondant à diverses estimations du taux d'erreur de classification fondées sur un modèle et le calcul de taux d'erreur réels correspondant aux seuils. La même série de 50 ensembles de données synthétiques était utilisée pour chaque ensemble de simulations. Il faut remarquer que l'ensemble C renferme 250,000 paires d'enregistrements, y compris 249,500 non-concordances vraies pour chaque essai de Monte Carlo. Afin de réduire le temps de calcul nécessaire pour effectuer les simulations, on n'a utilisé que 49,500 non-concordances vraies pour chaque essai. (On a effectué un essai à petite échelle afin de s'assurer que la réduction du nombre de non-concordances vraies avait un effet négligeable sur les probabilités d'accord estimées.) On a supprimé les non-concordances vraies contenues dans C en divisant les fichiers A et B en cinq blocs correspondants de taille 100 et en excluant les paires d'enregistrements dans lesquelles on trouvait des enregistrements provenant de blocs qui ne correspondaient pas.

Le système d'équations utilisé pour la méthode des moments a été résolu à l'aide d'une variation de la méthode de Newton, décrite en détail dans More et coll. (1980). Un logiciel fourni par IMSL (1987) a été utilisé. On a employé des probabilités d'accord de 0.9 pour les concordances vraies et de 0.1 pour les non-concordances vraies, pour tous les champs d'appariement, comme valeurs initiales pour la solution du système d'équations. La méthode ne semblait pas sensible aux valeurs initiales.

sur la mortalité. La fréquence de chaque vecteur de résultats parmi les concordances vraies est présentée au tableau 1. L'absence d'indépendance dans ces données est évidente. Bien que pour environ 88.3% des concordances vraies, il y ait accord pour le prénom, la probabilité d'un accord pour le prénom quand il y a désaccord pour l'initiale d'un deuxième prénom et accord pour le nom de famille et l'année de naissance n'est que de 381/1366, soit environ 27.9%. La valeur de la statistique du test du rapport des vraisemblances pour l'hypothèse d'indépendance est 3604. Cette valeur est extrêmement élevée par rapport à la distribution de référence chi carré avec 10 degrés de liberté. (Il faut remarquer qu'un degré de liberté est perdu parce que la fréquence pour la case (1,0,0,0) est zéro.)

Fréquences des résultats, ensemble de concordances vraies, données synthétiques

Résultat selon l'identificateur: 0 = Désaccord, 1 = Accord					Fréquence	
Initiale d'un deuxième prénom	Nom de famille	Année de naissance	Chiffre	Pourcen- tage		
0	0	0	0	0.03	7	0.03
0	0	0	1	0.12	33	0.12
0	0	1	0	0.45	125	0.45
0	0	1	1	0.85	985	3.54
0	1	0	0	0.02	5	0.02
0	1	0	1	0.14	39	0.14
0	1	1	0	0.73	202	0.73
0	1	1	1	6.65	1,848	6.65
1	0	0	0	0.0	0	0.0
1	0	0	1	0.05	13	0.05
1	0	1	0	0.18	50	0.18
1	0	1	1	1.37	381	1.37
1	1	0	0	0.16	44	0.16
1	1	0	1	1.62	451	1.62
1	1	1	0	6.30	1,751	6.30
1	1	1	1	78.65	21,860	78.65
Total					27,794	100

Pour chaque ensemble de données synthétiques, on a produit les enregistrements du fichier A en choisissant des identificateurs selon les fréquences relatives dans la Base canadienne de données sur la mortalité pour 1988. Afin de simplifier le processus de production des données, le choix des noms de famille était limité aux 100 noms de famille non francophones les plus courants et aux 100 noms de famille francophones les plus courants qui figuraient dans le fichier de 1988. Le choix des prénoms était

de la première variable d'appariement et

Ce modèle est conforme au modèle à variable latente général de Haberman (1979, p. 561). On doit imposer des restrictions additionnelles pour identifier et estimer les paramètres. Pour plus de simplicité, nous ne considérons que les modèles hiérarchiques. De plus, nous n'étudions que des modèles qui permettent à tous les effets non nuls d'interagir avec la variable latente.

non nuls d'interagir avec la variable latente.

à variable latente par les symboles $G(1)$, $G(2)$, ... les modèles log-linéaires pour les concordances vraies par $M(1)$, $M(2)$, ... et les modèles log-linéaires pour les non-concordances vraies par $U(1)$, $U(2)$, ... Dans le cas de quatre variables d'appariement, par exemple, le modèle $G(1)G(2)$, $G(3)$, $G(4)$ est un modèle à variable latente qui comprend un terme de niveau général, des effets principaux pour les quatre variables d'appariement et un terme pour l'interaction des variables d'appariement et un terme ainsi qu'un terme pour les effets principaux pour la variable latente ('interaction du terme de niveau général avec la variable latente), des termes pour l'interaction de chaque variable d'appariement et de la variable latente ainsi qu'un terme pour l'interaction des variables d'appariement.

On peut utiliser la méthode de pondération itérative de Haberman (1976) pour estimer des modèles à variable latente. La méthode d'estimation de Haberman consiste à soumettre à un balayage des tableaux qui renferment des chiffres estimés pour chaque résultat parmi les concordance vraies et les non-concordances vraies. Représentons les chiffres estimés pour le vecteur de résultats \bar{x} après i itérations de l'algorithme de Haberman par $C_{i,\bar{x}}$ et par $C_{i,\bar{x}}^0$ pour les concordances vraies et les non-concordances vraies, respectivement. On peut construire les valeurs initiales $C_{1,\bar{x}}^0$ et $C_{1,\bar{x}}^0$ à l'aide d'estimations des probabilités d'accord et de la proportion de concordances vraies obtenues quand on applique l'hypothèse de l'indépendance. Chaque itération de l'algorithme comporte une série d'opérations de balayage appliquées au tableau courant des concordances vraies ainsi que les opérations analogues sur le tableau courant des non-concordances vraies. À l'aide de la notation pour les modèles hiérarchiques présentée plus haut, on effectue un ensemble d'opérations de balayage pour chacun des termes d'interaction qui définissent le modèle. Pour quatre variables d'appariement ainsi que le

Le traitement de l'algorithme prend fin quand les variations entre les chiffres estimés pour des itérations consécutives sont inférieures à une tolérance donnée. Haberman (1976) fait remarquer qu'il se peut que l'algorithme de pondération itérative converge vers un maximum local de la fonction de vraisemblance plutôt que vers l'estimation obtenue à l'aide de la méthode du maximum de vraisemblance. Des expériences réalisées avec des valeurs initiales différentes qui utilisent des ensembles de données employés dans l'évaluation dont on traite à la section 5 n'ont permis d'obtenir aucun exemple de ce problème.

On présente, dans cette section, les résultats de comparaisons des méthodes d'estimation décrites dans les sections 3 et 4. Les comparaisons comprenaient l'application de chaque méthode à une série d'ensembles de données synthétiques produits à l'aide de méthodes de Monte Carlo. On a employé des enregistrements de données synthétiques renfermant quatre identificateurs de personne (nom de famille, initiale d'un deuxième prénom, prénom, date de naissance). Les renseignements sur les valeurs possibles de chaque identificateur ainsi que leurs fréquences relatives, ont été tirés de la Base canadienne de données sur la mortalité pour 1988. Cette base de données, qui est souvent utilisée dans des applications de couplage d'enregistrements dans le domaine de la santé, renferme un enregistrement distinct pour chaque mortalité.

L'hypothèse d'indépendance n'était pas respectée parmi les concordances vraies dans chaque ensemble de données synthétiques. Pendant la production des données, on a utilisé des renseignements, sur la fréquence des vecteurs de résultats pour les concordances vraies, obtenus à partir de divers projets de couplage d'enregistrements réalisés par le Centre canadien d'information sur la santé de Statistique Canada. La majorité des projets comportaient l'appariement d'un fichier de cohorte à la Base canadienne de données

$${}^{1/8}S_{\bar{3}\bar{3}}\bar{X}A = \bar{X}^{\dagger}{}_{1-1/2}\bar{3} / \bar{X}^{\dagger}{}_{1-1/2}\bar{3} = \bar{X}^{\dagger}{}_{1-1/2}\bar{3}$$

modèle $G(1)G(2)$, $G(3)G(4)$, on effectue deux ensembles d'opérations de balayage, une pour l'interaction $G(1)G(2)$ et la seconde pour l'interaction $G(3)G(4)$. Pour chaque interaction, une opération de balayage est effectuée pour chaque niveau de la variable de classification correspondante. Représentons par S_{gi}^l l'ensemble de vecteurs de résultats au niveau l du terme g . L'opération de balayage appliquée au tableau des concordances vraies lors de l'itération l pour le niveau l du terme g comporte le calcul de

La méthode exige des estimations initiales des probabilités d'accord pour les concordances vraies et pour les non-concordances vraies. Pour les concordances vraies, on doit utiliser des estimations au jugé basées sur l'expérience acquise. Pour obtenir des estimations initiales des probabilités d'accord parmi les paires d'enregistrements qui correspondent à des non-concordances vraies, on suppose habituellement que ces probabilités sont égales aux probabilités d'accord parmi les paires d'enregistrements choisies de façon aléatoire, nommément que:

$$u_k = P(x_k = 1), \quad k = 1, 2, \dots, K.$$

Supposons que $J(k)$ valeurs différentes apparaissent, pour le champ de données k , dans le fichier A et/ou dans le fichier B. Désignons les fréquences de ces valeurs dans le fichier A par $f_{k1}, f_{k2}, \dots, f_{kj(k)}$ et désignons les fréquences pour le fichier B par $g_{k1}, g_{k2}, \dots, g_{kj(k)}$. Pour une valeur particulière, un des chiffres, mais non les deux, peut être nul. L'estimation initiale de u_k est

$$u_k^0 = \sum_{j=1}^{J(k)} (f_{kj} g_{kj}) / N. \quad (8)$$

Compte tenu de ces estimations de probabilités, des ensembles initiaux de concordances et de non-concordances, désignés par M^0 et par U^0 , respectivement, sont obtenus à l'aide d'une règle de décision

$$\begin{aligned} j \in M^0 & \quad \text{si} \quad \omega_j > \tau_1^0, \\ j \in U^0 & \quad \text{si} \quad \omega_j > \tau_2^0. \end{aligned}$$

Puis, on utilise les chiffres des fréquences parmi les paires d'enregistrements dans les ensembles M^0 et U^0 comme nouvelles estimations des probabilités d'accord. Ces estimations sont employées pour obtenir de nouveaux ensembles de concordances et de non-concordances et le processus itératif est repris jusqu'à ce que des estimations consécutives des probabilités d'accord soient suffisamment rapprochées.

Dans la majorité des applications, l'hypothèse que la probabilité d'accord parmi les paires d'enregistrements qui correspondent à des concordances vraies est égale à la probabilité d'accord parmi toutes les paires d'enregistrements est justifiée et l'itération ne mène pas à des modifications importantes dans les estimations des probabilités d'accord pour les non-concordances vraies. Toutefois, il arrive souvent que la première itération produise des variations considérables dans les estimations de la probabilité d'accord pour les concordances vraies. Généralement, il n'y a pas de variations importantes lors de la deuxième itération. Il faudrait remarquer que les propriétés statistiques de la méthode itérative ne sont pas connues avec précision. En pratique, le rendement de la méthode dépendra du choix des seuils initiaux τ_1^0, τ_2^0 . Ces seuils sont généralement

choisis de façon subjective. Les simulations mentionnées dans la section 5 fournissent des renseignements à propos des effets des divers seuils initiaux.

4. ASSOUPLISSEMENT DE L'HYPOTHÈSE DE L'INDÉPENDANCE - ESTIMATION À L'AIDE DE LA PONDERATION ITÉRATIVE

On peut utiliser des méthodes d'estimation pour les modèles à variable latente afin d'estimer les probabilités d'accord quand on effectue le paramétrage, en fonction d'effets log-linéaires, de la dépendance entre les résultats de comparaisons pour différents champs d'appariement. Winkler (1989) et Thibaudreau (1989) ont estimé les probabilités d'accord avec des modèles log-linéaires qui comprennent tous les termes d'interaction jusqu'au troisième ou quatrième ordre afin de paramétrer les dépendances. La formulation présentée ici facilite l'utilisation de modèles log-linéaires, y compris certaines interactions. On peut considérer l'état par rapport à la concordance comme une variable latente avec deux niveaux (concordance vraie et non-concordance vraie). Représentons par $c_{0,\bar{x}}$ et $c_{1,\bar{x}}$ le nombre de non-concordances vraies et de concordances vraies, respectivement, avec vecteur de résultats \bar{x} dans une application de couplage d'enregistrements pour laquelle on utilise K variables d'appariement. Bien entendu, ces chiffres ne peuvent être observés puisque la valeur de la variable latente pour chaque paire d'enregistrements est inconnue. On observe plutôt $c_{\bar{x}} = c_{0,\bar{x}} + c_{1,\bar{x}}$. À l'aide du paramétrage de la dépendance réalisé au moyen d'effets log-linéaires et d'un modèle saturé pour les concordances vraies, nous pouvons poser

$$\begin{aligned} \log(c_{1,\bar{x}}/Np) &= M(0) + M(1)x_1 + M(2)x_2 + \dots \\ &+ M(K)x_K + M(1)M(2)x_1x_2 + \dots \\ &+ M(K) - 1)M(K)x_K - 1)x_K + \dots \\ &+ M(1)M(2) \dots M(K)x_1x_2, \dots, x_K, \end{aligned}$$

avec les restrictions habituelles. Nous disposons d'une expression semblable pour les non-concordances vraies. Le modèle à variable latente correspondant à ces modèles log-linéaires saturés est:

$$\begin{aligned} \log(c_{s,\bar{x}}/w_s) &= G(0) + Z_s + G(1)x_1 + \dots \\ &+ G(K)x_K + ZG(1)_{s,x_1} + \dots + ZG(K)_{s,x_K} \\ &+ \dots + G(1)G(2) \dots G(K)_{x_1x_2, \dots, x_K} \\ &+ ZG(1)G(2) \dots G(K)_{s,x_1x_2, \dots, x_K}, \end{aligned}$$

où la valeur de l'indice s est zéro pour les non-concordances vraies et un pour les concordances vraies,

$$\log(P(\bar{x}|M)) = M(0) + M(1)^{x_1} + M(2)^{x_2} + \dots + M(K)^{x_K} + M(1)M(2)^{x_1, x_2} + \dots + M(K-1)M(K)^{x_{K-1}, x_K} + \dots + M(1)M(2) \dots M(K)^{x_1, x_2, \dots, x_K}, \quad (6)$$

avec les restrictions habituelles

$$\sum_{x_J} M(J)^{x_J} = 0, \quad J = 1, 2, \dots, K, \\ \sum_{x_{J_1}, x_{J_2}} M(J_1)M(J_2)^{x_{J_1}, x_{J_2}} = \sum_{x_{J_2}} M(J_1)M(J_2)^{x_{J_1}, x_{J_2}} = 0, \\ \forall J_1, J_2, \dots, \text{etc.},$$

ainsi que la restriction

$$\sum_{\bar{x}} P(\bar{x}|M) = 1.$$

Le modèle saturé pour $P(\bar{x}|U)$ est analogue.

Si l'on utilise des modèles log-linéaires saturés pour $P(\bar{x}|M)$ et pour $P(\bar{x}|U)$, la fonction de densité comprend $2^{K+1} - 1$ paramètres inconnus. On ne peut identifier tous ces paramètres quand on ne dispose pas de renseignements auxiliaires. Afin d'obtenir un modèle qui peut être identifié et pour simplifier le problème d'estimation, on fait souvent l'hypothèse que les résultats des comparaisons pour différents champs de données sont indépendants. Quand on suppose qu'il y a indépendance, on désigne les probabilités d'accord parmi les paires d'enregistrements qui sont des concordances vraies et des non-concordances vraies, respectivement, par

$$m_k = P(x_k = 1 | M), \quad k = 1, 2, \dots, K,$$

$$u_k = P(x_k = 1 | U), \quad k = 1, 2, \dots, K.$$

Les probabilités de résultats peuvent être écrites sous la forme:

$$P(\bar{x}|M) = \prod_{k=1}^K m_k^{x_k} (1 - m_k)^{(1-x_k)},$$

$$P(\bar{x}|U) = \prod_{k=1}^K u_k^{x_k} (1 - u_k)^{(1-x_k)}.$$

Ce modèle comprend $2 \cdot K + 1$ paramètres inconnus, normalement (\bar{m}, \bar{n}, p) , où $\bar{m} = (m_1, m_2, \dots, m_K)$, $\bar{n} = (u_1, u_2, \dots, u_K)$. Il y a, bien entendu, un certain nombre de modèles intermédiaires entre le modèle saturé

et le modèle où l'on fait appel à l'indépendance. Des méthodes qui peuvent être utilisées pour estimer le modèle qui fait appel à l'indépendance sont décrites dans la section 3. On traite de l'estimation de modèles intermédiaires dans la section 4.

3. ESTIMATION FONDÉE SUR L'HYPOTHÈSE DE L'INDÉPENDANCE

3.1 Méthode des moments

On peut employer un estimateur de $P(\bar{x}|M)$ et de $P(\bar{x}|U)$ obtenu à l'aide de la méthode des moments quand il y a indépendance. L'estimateur est basé sur un système de $2 \cdot K + 1$ équations qui fournissent des expressions pour des moments fonctionnellement indépendants de \bar{x} en fonction des paramètres. Les équations sont les suivantes:

$$E\left(\prod_{k=1}^K x_k\right) = p^N N^{\prod_{k=1}^K m_k + (1-p)} \prod_{k=1}^K u_k,$$

$$i = 1, 2, \dots, K$$

$$E(x_i) = p^N m_i + (1-p) N u_i, \quad i = 1, 2, \dots, K, \quad (7)$$

Pour obtenir des estimations des paramètres à l'aide de la méthode des moments, il faut résoudre les équations une fois que les valeurs espérées ont été remplacées par des moyennes calculées à l'aide de paires d'enregistrements dans C. Le système d'équations pour $K = 3$ a été présenté par Fellegi et Sunter, qui ont aussi calculé une solution en forme analytique fermée qui existe si certaines conditions modérées sont satisfaites. Leur article comprenait une mise en garde pour ce qui est de l'utilisation de la méthode s'il n'y a pas d'indépendance. Pour $K > 3$, on ne dispose pas d'une solution en forme analytique fermée, mais on peut employer les méthodes numériques courantes. Les estimations de paramètres obtenues à l'aide de la méthode des moments sont statistiquement convergentes si l'hypothèse d'indépendance est vérifiée.

3.2 Méthode itérative

La méthode itérative a été élaborée par des personnes qui effectuent des couplages d'enregistrements. Bien que la méthode ne soit pas basée sur la distribution de probabilité du vecteur de résultats, elle fait appel à l'hypothèse d'indépendance. L'application de la méthode itérative est décrite par plusieurs auteurs, y compris Newcombe (1988). Le logiciel de couplage d'enregistrements de Statistique Canada, CANLINK, est conçu de façon à faciliter l'utilisation de la méthode itérative.

Pour nos fins, nous ne tenons compte que des résultats de l'accord et du désaccord. Dans le cas de K champs d'appariement, nous définissons le vecteur de résultats $\bar{x}' = (x_1', x_2', \dots, x_K')$ pour la paire d'enregistrements j . Nous avons $x_k' = 1$ s'il y a concordance pour la paire d'enregistrements j à propos du champ de données k et $x_k' = 0$ s'il y a non-concordance pour la paire d'enregistrements j à propos du champ de données k . Newcombe et coll. (1959) ont proposé l'idée que les décisions relatives au fait qu'une paire d'enregistrements représente ou non la même entité devraient être basées sur le rapport

$$R(\bar{x}) = P(\bar{x} | M) / P(\bar{x} | U), \quad (1)$$

où $\bar{x} = (x_1, x_2, \dots, x_K)$ est le vecteur de résultats généraux, $P(\bar{x} | M)$ est la probabilité que les comparaisons pour une paire d'enregistrements où il y a concordance vraie produiront le vecteur de résultats \bar{x} , et $P(\bar{x} | U)$ est la probabilité de \bar{x} pour une paire d'enregistrements pour laquelle il y a non-concordance vraie. L'optimalité des méthodes de couplage d'enregistrements qui font appel à ce rapport a été démontrée par Fellegi et Sunter.

Dans le modèle de Fellegi-Sunter, une règle d'appariement attribue une probabilité pour chaque décision de classification (A_1, A_2 et A_3) à chaque vecteur de résultats. La fonction de décision correspondant au vecteur de résultats \bar{x} est $d(\bar{x}) = (P(A_1 | \bar{x}), P(A_2 | \bar{x}), P(A_3 | \bar{x}))$. Des taux d'erreur de classification acceptables pour les non-concordances vraies et les concordances vraies sont précisés avant que le couplage ne soit effectué. Nous désignons ces taux d'erreur précisés à l'avance par μ et λ respectivement. Fellegi et Sunter définissent, parmi la classe de règles d'appariement d'enregistrements qui satisfont aux relations $P(A_1 | U) \leq \mu$ et $P(A_3 | M) \leq \lambda$ pour des valeurs fixes de μ et de λ , la règle d'appariement optimale comme étant la règle qui minimise $P(A_2)$, la probabilité qu'une paire d'enregistrements sera classée comme un cas indéterminé. La règle optimale a la forme

$$\begin{aligned} d(\bar{x}') &= (1, 0, 0) & \text{si } \omega' > \tau_1 \\ d(\bar{x}') &= (P_\mu, 1 - P_\mu, 0) & \text{si } \omega' = \tau_1 \\ d(\bar{x}') &= (0, 1, 0) & \text{si } \tau_2 < \omega' < \tau_1 \\ d(\bar{x}') &= (0, 1 - P_\lambda, P_\lambda) & \text{si } \omega' = \tau_2 \\ d(\bar{x}') &= (0, 0, 1) & \text{si } \omega' < \tau_2 \end{aligned} \quad (2)$$

où $\tau_1 \geq \tau_2$, le "poids" ω' est défini comme $\omega' = \log(R(\bar{x}'))$ et P_μ et P_λ sont des constantes positives sur l'intervalle $[0, 1]$. (Pour tous les détails, consulter Fellegi et Sunter (1969)). Pour déterminer τ_1 et τ_2 , il faut estimer les taux d'erreur de classification correspondant à divers choix pour ces valeurs seuil, ce qui souligne l'importance d'une estimation précise des taux d'erreur de classification dans le modèle de Fellegi-Sunter.

$$\hat{\lambda} = \sum_{\bar{x} \in L(\tau_2)} P(\bar{x} | M) + P_\lambda \sum_{\bar{x} \in Q(\tau_2)} P(\bar{x} | M) \quad (3)$$

où $L(\tau_2) = \{\bar{x}; \log(R(\bar{x})) < \tau_2\}$ et $Q(\tau_2) = \{\bar{x}; \log(R(\bar{x})) = \tau_2\}$.

L'estimation du taux d'erreur de classification pour les non-concordances vraies fondée sur un modèle est

$$\hat{\mu} = \sum_{\bar{x} \in G(\tau_1)} P(\bar{x} | U) + P_\mu \sum_{\bar{x} \in Q(\tau_1)} P(\bar{x} | U) \quad (4)$$

où $G(\tau_1) = \{\bar{x}; \log(R(\bar{x})) > \tau_1\}$ et $Q(\tau_1) = \{\bar{x}; \log(R(\bar{x})) = \tau_1\}$.

2.2 Un modèle pour les probabilités de résultats

Pour calculer les estimations du taux d'erreur de classification fondées sur un modèle, il faut estimer $P(\bar{x} | M)$ et $P(\bar{x} | U)$, pour chacune des 2^K valeurs possibles de \bar{x} . La densité de probabilité pour \bar{x} est une combinaison de deux densités de probabilité données par

$$f(\bar{x}) = pP(\bar{x} | M) + (1 - p)P(\bar{x} | U), \quad (5)$$

où p est la probabilité qu'il y a concordance vraie pour une paire d'enregistrements choisis au hasard. Les probabilités de résultats dépendent de la distribution de fréquences des identificateurs pour des entités représentées dans les fichiers A et B , ainsi que des probabilités qu'il y a introduction d'erreurs quand les identificateurs sont enregistrés dans les fichiers. Fellegi et Sunter (1969, pp. 1192-1194) décrivent, pour estimer les probabilités d'accord, une méthode qui fait appel à leur définition en fonction des distributions de fréquences et des probabilités d'erreur. Ils recommandent d'utiliser la méthode quand on dispose d'informations préalables.

Dans le présent article, nous considérons des situations où les données dans les fichiers A et B , ainsi que les vecteurs de résultats \bar{x}' , $j = 1, 2, \dots, N$, représentent les seuls renseignements disponibles pour l'estimation des probabilités de résultats. Une structure log-linéaire pour les probabilités de résultats constitue le paramétrage le plus général. Le modèle log-linéaire saturé pour les probabilités de résultats dans le cas des concordances vraies est

fournissent pas une preuve suffisante pour justifier la classification de la paire comme un lien ou un non-lien à des niveaux d'erreur inférieurs ou égaux aux niveaux précisés. Il faut estimer avec précision les taux d'erreur de classification associées à diverses règles de décision afin de déterminer une règle appropriée. Le taux d'erreur de classification pour les non-concordances vraies est $P(A_1 | U)$. Le taux d'erreur de classification pour les concordances vraies est $P(A_2 | M)$.

On peut obtenir des estimations des taux d'erreur de classification en choisissant un échantillon de paires d'enregistrements de l'ensemble C et en déterminant manuellement l'état véritable vis-à-vis de la concordance des paires échantillonnées. Des applications de cette méthode sont décrites dans Bartlett et coll. (1993). L'échantillonnage peut être à la fois coûteux et peu facile à réaliser, particulièrement quand on doit effectuer le même couplage pour un certain nombre de paires de fichiers, chacune avec des caractéristiques légèrement différentes. Bellin et Rubin (1991) décrivent une autre méthode d'estimation des taux d'erreur pour lesquelles il faut disposer de l'état véritable de la concordance pour des paires d'enregistrements dans une étude-pilote. Par opposition à la méthode d'échantillonnage simple, la méthode de Bellin et Rubin fournit un modèle pour l'application de renseignements, obtenus à l'aide de l'étude-pilote, à des couplages plus importants dans lesquels des données semblables sont utilisées.

Le modèle de Fellegi-Sunter fournit une méthode pour calculer des estimations des taux d'erreur à l'aide d'estimations des probabilités qu'il y aura concordance entre des paires d'enregistrements pour diverses combinaisons de champs de données. Le calcul de ces estimations du taux d'erreur obtenues à l'aide de modèles est simple et la détermination manuelle de l'état de paires d'enregistrements par rapport à la concordance vraie n'est pas nécessaire. Cependant, ces estimations ont souvent des propriétés qui laissent à désirer dans des applications. Voir, par exemple, Bellin (1990). Dans le présent article, on démontre que les propriétés des estimations du taux d'erreur obtenues à l'aide de modèles peuvent être améliorées en faisant une estimation soignée des probabilités d'accord.

Trois méthodes d'estimation qui peuvent être employées sont évaluées. Les façons de procéder décrites n'utilisent que les renseignements dans les fichiers A et B . Elles n'emploient pas de renseignements auxiliaires. Les estimations des taux d'erreur obtenues à l'aide de modèles pour chacune des méthodes évaluées sont comparées aux taux d'erreur réels à l'aide de données synthétiques qui incorporent des caractéristiques importantes de données tirées d'applications du couplage d'enregistrements dans le domaine de la santé que de renseignements obtenus à partir d'une application réelle du couplage d'enregistrements. Voici la structure du présent article. La section 2 contient des détails sur la méthode d'estimation de l'erreur de classification à l'aide d'un modèle présentée par Fellegi et Sunter. Le modèle utilisé pour les probabilités d'accord qui forme la base de la discussion ultérieure des méthodes d'estimation est aussi précisé. Deux méthodes d'estimation

qui sont fondées sur une hypothèse d'indépendance importante sont décrites dans la section 3. On traite d'une troisième méthode pour laquelle l'indépendance n'est pas exigée dans la section 4. Les résultats des comparaisons des trois méthodes à l'aide de données synthétiques sont présentés dans la section 5. Les résultats du travail d'évaluation effectué avec des renseignements tirés d'une application réelle sont décrits dans la section 6. La section 7 renferme des conclusions.

2. CONCEPTS THÉORIQUES

Nous résumons, dans la présente section, les aspects pertinents de la théorie du couplage d'enregistrements élaborée par Fellegi et Sunter (1969). Dans le modèle de Fellegi-Sunter, les estimations des taux d'erreur de classification sont calculées à l'aide d'estimations des probabilités d'accord pour diverses combinaisons de champs de données. Les applications de la théorie de Fellegi et Sunter supposent habituellement l'hypothèse que la probabilité qu'il y aura accord, pour une paire d'enregistrements, au niveau d'un champ de données particulier est indépendante des résultats des comparaisons pour d'autres champs. La théorie est néanmoins très flexible, pouvant tenir compte de tous les types de dépendance entre les résultats de comparaisons pour différents champs de données. Un paramétrage de la dépendance en fonction d'effets log-linéaires est présenté.

2.1 Estimation des taux d'erreur de classification fondée sur un modèle

Pour obtenir des renseignements portant sur la classification d'une paire d'enregistrements comme un lien (A_1), un non-lien (A_2) ou un cas indéterminé (A_3), on compare des champs de données renfermant des renseignements d'identification. Dans une application comportant des enregistrements relatifs à des personnes, on pourrait faire des comparaisons distinctes des noms de famille, des prénoms et des dates de naissance. Le résultat d'une comparaison est un code numérique représentant un énoncé du genre: "les noms concordent", "les noms ne concordent pas", "un nom manque dans un des fichiers ou dans les deux", "les noms ne concordent pas mais leurs deux premiers caractères concordent". Les codes de résultat utilisés dans les travaux d'application diffèrent selon les applications et selon les comparaisons dans la même application. Le plus petit nombre de codes de résultat qui peut être utilisé pour toute comparaison est deux, ce qui correspond à l'accord et au désaccord. Dans les applications, il faut habituellement disposer d'un code de résultat correspondant à "manquant dans un des deux fichiers ou dans les deux". Le résultat de l'accord peut être remplacé par un certain nombre de résultats ayant une valeur particulière (telle que "les noms concordent et ils sont tous deux Georges"). Certains désaccords peuvent être codés comme des accords partiels (tels que "les noms ne concordent pas mais leurs deux premiers caractères concordent").

Estimation modeliste des taux d'erreur liés au couplage d'enregistrements

J.B. ARMSTRONG et J.E. MAYDA¹

RÉSUMÉ

On appelle couplage d'enregistrements l'appariement d'enregistrements contenant des données sur des particuliers, pratiques, comportent la classification de paires d'enregistrements, comme constituant des liens ou des non-liens, à l'aide d'une procédure automatisée basée sur le modèle théorique présenté par Fellegi et Sunter (1969). L'estimation des taux d'erreur de classification constitue un problème important. Fellegi et Sunter présentent une méthode, afin de calculer des estimations des taux d'erreur de classification, qui découle directement du couplage. Ces estimations faites à l'aide de modèles sont plus faciles à produire que celles obtenues par appariement manuel d'échantillons, méthode généralement utilisée en pratique. Les propriétés des estimations du taux d'erreur de classification fondées sur un modèle, obtenues au moyen de trois estimateurs de paramètre de modèle, sont comparées.

MOTS CLÉS: Modèle mixte; modèle à variable latente; pondération itérative.

1. INTRODUCTION

Dans de nombreuses applications statistiques, on utilise des fichiers informatiques renfermant des renseignements sur des particuliers, sur des entreprises et sur des logements. Il faut souvent effectuer le couplage d'enregistrements qui se rapportent à la même entité. L'opération, qui consiste à coupler des enregistrements portant sur la même entité s'appelle appariement exact. Si l'on a attribué un identificateur unique à tous les enregistrements utilisés dans une application, l'appariement exact ne pose aucun problème. Les méthodes de couplage d'enregistrements s'attaquent au problème de l'appariement exact quand on ne dispose pas d'un identificateur unique. Dans ce cas, chaque enregistrement inclut généralement un certain nombre de champs de données renfermant des renseignements d'identification qui pourraient être utilisés pour effectuer l'appariement. Les problèmes rencontrés au cours de l'appariement sont dus à des erreurs dans ces données ou au fait que la même valeur dans un champ particulier est valable pour plus d'une entité.

Les applications du couplage d'enregistrements comprennent l'élimination des doubles comptes dans des listes de logements ou d'entreprises obtenues de diverses sources afin de créer des bases de sondage. De plus, le couplage d'enregistrements est largement utilisé dans des applications portant sur la santé et l'épidémiologie. Le travail dans ce domaine comporte généralement l'appariement d'enregistrements renfermant des renseignements sur des particuliers dans des cohortes d'industries ou de professions avec des enregistrements renfermant des renseignements sur la maladie ou sur le décès de particuliers. Par exemple, dans Fair, Newcombe et Lalonde (1988), on traite des

méthodes de couplage d'enregistrements utilisées dans des études de suivi portant sur des personnes exposées à la radiation.

Le problème du couplage d'enregistrements peut être formulé à l'aide de deux fichiers de données qui correspondent à deux populations. Chaque fichier peut renfermer soit de l'information sur toutes les entités dans la population correspondante, soit de l'information pour un échantillon aléatoire d'entités. Le fichier A contient N_A enregistrements et le fichier B en renferme N_B . L'ensemble de paires d'enregistrements formées comme produit croisé de A et de B, est représenté par $C = \{(a,b); a \in A, b \in B\}$. C contient $N = N_A \cdot N_B$ paires d'enregistrements. L'objectif du couplage d'enregistrements est de partager l'ensemble C en deux ensembles disjoints, l'ensemble des concordances vraies, représenté par M et l'ensemble des non-concordances vraies, U.

De nombreux travaux d'application sont basés sur le modèle théorique présenté par Fellegi et Sunter (1969). Pour chaque paire d'enregistrements, on prend une décision afin de déterminer si les enregistrements se rapportent ou non à la même entité, après avoir examiné les données enregistrées dans les fichiers A et B. Les décisions possibles sont les suivantes: lien (A_1), non-lien (A_2) et cas indéterminé (A_3). Il y a deux types d'erreurs. Premièrement, la décision A_1 peut être prise pour une paire d'enregistrements qui fait partie de U, l'ensemble des non-concordances vraies. Deuxièmement, la décision A_3 peut être prise pour une paire d'enregistrements qui fait partie de M, l'ensemble des concordances vraies. Des niveaux acceptables d'erreur de classification sont précisés avant le couplage des fichiers. Une paire d'enregistrements est classée comme un cas indéterminé si les données ne

Nous avons aussi l'intention, dans des études futures, de considérer des modèles multidimensionnels avec ordonnée à l'origine pour l'estimation de l'erreur systématique de mesure. Comme les critères d'évaluation de la méthode d'auto-amorçage élaborés dans cet article sont de nature générale, il n'est pas nécessaire de modifier la méthode d'évaluation pour ajouter des variables dans les modèles d'estimation. En outre, nous prévoyons améliorer et évaluer les hypothèses de modèle et les méthodes de traitement des valeurs aberrantes dans un article ultérieur. Finalement, il nous faut étudier l'effet de la non-vérification des hypothèses de modèle sur l'estimation, en particulier l'hypothèse selon laquelle les données de réinterview ne sont entachées d'aucune erreur. Comme le montre Fuller (1991), si les données de la réinterview sont faillibles mais non biaisées, la variance des valeurs prédites s'accroît mais les valeurs prédites demeurent non biaisées. Par conséquent, on pourrait étudier, suivant ces hypothèses, la précision relative des divers estimateurs de l'erreur systématique de mesure afin de déterminer la robustesse de la méthode de prédiction modeliste.

REMERCIEMENTS

Les auteurs remercient Manuel Cardenas pour sa précieuse aide ainsi que les arbitres pour leurs remarques pertinentes.

BIBLIOGRAPHIE

- COCHRAN, W. (1977). *Sampling Techniques*. New York: John Wiley and Sons.
- EFFRON, B., et GONG, G. (1983). A leisurely look at the bootstrap, the jackknife, and cross-validation. *The American Statistician*, 31, 36-48.
- FORSMAN, G., et SCHREINER, I. (1991). The design and analysis of reinterview: an overview. Dans *Measurement Errors in Surveys*. (Eds. P. Biemer, et coll.). New York: John Wiley and Sons.
- FULLER, W.A. (1991). Regression estimation in the presence of measurement error. Dans *Measurement Errors in Surveys*. (Eds. P.P. Biemer, et coll.). New York: John Wiley and Sons, 617-636.
- HANSEN, M., MADOW W., et TEPPING, B. (1983). An evaluation of model-dependent and probability sampling inferences in sample surveys. *Journal of the American Statistical Association*, 78, 776-793.
- RAO, J.N.K., et WU, C. (1988). Resampling inference with complex survey data. *Journal of the American Statistical Association*, 83, 231-241.
- ROYALL, R., et HERSON, J. (1973). Robust estimation in finite populations I. *Journal of the American Statistical Association*, 68, 880-893.
- ROYALL, R., et CUMBERLAND, W. (1978). Variance estimation in finite population sampling. *Journal of the American Statistical Association*, 73, 351-361.
- ROYALL, R., et CUMBERLAND, W. (1981). The finite-population linear regression estimator and estimators of its variance – an empirical study. *Journal of the American Statistical Association*, 76, 924-930.
- SÄRNDAAL, C.-E., SWENSSON, B., et WRETMAN, J. (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.
- BIEMER, P., et STOKES, L. (1991). Approaches to the modeling of measurement errors. Dans *Measurement Errors in Surveys*. (Eds. P. Biemer, et coll.). New York: John Wiley and Sons.
- BICKEL, P., et FREEDMAN, D.A. (1984). Asymptotic normality and the bootstrap in stratified sampling. *The Annals of Statistics*, 12, 470-482.

3.3 Sommaire des résultats

Le Tableau 2 présente un sommaire des résultats de notre étude. La première colonne de chiffres indique la valeur connue de $B^* = E(Y_i^* - \mu_i^*)$, qui est le paramètre de biais pour la pseudo-population U^* . Les autres colonnes de chiffres contiennent les valeurs des estimateurs, avec les erreurs types entre parenthèses (e.l. $(\hat{\theta}) = \sqrt{\text{var}_{BSS}(\hat{\theta})}$). Les quatre dernières lignes du tableau indiquent, respectivement,

a) le nombre de postes (sur 10) pour lesquels la valeur B^* est incluse dans un intervalle de confiance à 95%,

b) le coefficient de variation (C.V.) moyen,

c) la moyenne de la racine carrée de \widehat{EQM}^* et

d) la valeur moyenne du biais relatif en valeur absolue.

Un trait saillant de ces résultats est leur forte disparité d'un estimateur à l'autre pour toutes les caractéristiques étudiées, en particulier les stocks de blé (toutes variétés). Dans ce dernier cas, les estimations varient de -94.2 à 103.2. Le Tableau 2 indique aussi, par le symbole ‡, si le paramètre B^* est contenu dans un intervalle de confiance à 95%, notamment $[\hat{\theta} - 2 \text{ e.l.}(\hat{\theta}), \hat{\theta} + 2 \text{ e.l.}(\hat{\theta})]$. Le meilleur estimateur à ce titre est B_{SSW} , pour lequel nous avons des intervalles de confiance qui contiennent B^* dans huit cas sur dix. L'estimateur B_{2st} vient au deuxième rang, avec six postes sur dix, et B_M , au troisième rang, avec cinq postes sur dix. L'estimateur par quotient classique et sa version robuste donnent les pires résultats, la valeur de B^* étant contenue dans l'intervalle de confiance dans un cas seulement.

Si on examine ces estimateurs du point de vue de l'erreur quadratique moyenne, on fait des constatations différentes. En effet, B_M s'impose comme l'estimateur pour lequel la valeur moyenne de la racine carrée de l'EQM est la plus faible. Cependant, les estimateurs B_{SSW} et B_{2st} ne sont pas loin derrière. En outre, B_{SSW} est l'estimateur pour lequel la valeur moyenne du biais relatif en valeur absolue est la plus faible. Cet estimateur a produit un biais appréciable dans seulement deux cas sur dix. Les résultats nous permettent donc de croire que B_{SSW} est le meilleur estimateur, si l'on se sert du rendement global comme critère d'évaluation.

4. CONCLUSIONS ET RECOMMANDATIONS

Dans cet article, nous avons élaboré une méthode générale pour construire et évaluer des estimateurs de prédiction modéliste, pondérés ou non, de l'erreur systématique de mesure pour des plans d'échantillonnage aléatoire à deux phases stratifié. Les estimateurs proposés renferment de l'information sur les observations y (information tirée de l'échantillon de la première phase) et sur une variable auxiliaire x . Nous avons aussi considéré et évalué des versions robustes de ces estimateurs. Le but ultime de l'estimation par la prédiction modéliste est de définir des estimateurs qui fount un usage "optimal" des données (y, μ, x) . La

La prédiction modéliste permet d'extraire des réinterviews le maximum d'information sur le biais dans les réponses. Les estimateurs de prédiction modéliste seront donc généralement plus efficaces que l'estimateur TEN classique. De plus, la prédiction modéliste offre aussi un moyen d'étendre les estimations du biais à des domaines qui n'ont pas été sondés. Dans les enquêtes de réinterview du NASS par exemple, pour des raisons de commodité et d'économie, l'échantillon était constitué uniquement d'unités qui avaient participé à l'enquête initiale par ITAO. Or, en utilisant des modèles de prédiction qui sont des fonctions des réponses originales et des caractéristiques géographiques et démographiques locales, il serait possible de prédire l'erreur systématique de mesure pour les unités d'enquête qui participent autrement que par ITAO en se fondant sur les caractéristiques locales de ces unités, ce qui serait un genre d'estimation "synthétique". Bien que cette application de l'estimation basée sur un modèle n'ait pas été étudiée dans cet article, elle est un prolongement naturel de la méthode exposée et sera évaluée dans une étude ultérieure.

3.2 Comparaison des estimateurs de M et de B

En nous servant des données de l'enquête agricole de décembre 1990 et des données correspondantes recueillies à la réinterview, nous avons comparé les estimateurs éla-
estimations de l'erreur type et de l'erreur quadratique
moyenne à l'aide de la méthode d'auto-amorçage de
Bickel-Freedman décrite dans la section 2.4 ($\bar{Q} = 300$
échantillons). Le Tableau 2 contient les résultats relatifs
à six estimateurs, soit B_{2st} , l'estimateur de différence clas-
sique; B_{x2stR} , l'estimateur par quotient pondéré; B_{x2stR} ,
la version robuste (suppression de valeurs aberrantes)
de B_{x2stR} ; l'estimateur de Särndal, Swensson et
Wretman; B_M , l'estimateur à modèle non pondéré; et B_M ,
la version robuste (suppression de valeurs aberrantes)
de B_M .

représentant un certain nombre de questions des enquêtes
initiales. Afin de réduire au maximum les erreurs de
mémoire, on a fait les réinterviews dans les 10 jours qui
suivaient l'interview initiale. On s'efforçait de concilier les
différences entre les réponses données à l'EA et celles don-
nées à la réinterview pour déterminer la valeur "vraie".
Beaucoup d'efforts ont été consacrés à l'élaboration de
procédures, à la formation et à la supervision du processus
de réinterview pour faire en sorte que les réponses défini-
tives soient toutes justes. Dans la grande majorité des cas,
le libellé des questions posées dans la réinterview était le
même que dans l'enquête d'origine. Les intervieweurs ten-
taient de communiquer avec la personne la plus avertie
pour s'assurer de l'exactitude des réponses données.
Dans ce rapport, nous n'analysons que les données de
1990. Le Tableau 1 donne la taille des échantillons de réin-
terview utilisés pour cette étude.

Tableau 2

Comparaison des estimateurs étudiés, par rapport à B^* , la valeur du biais pour la pseudo-population †

Caractéristique	B^*	B_{2st}	B_{x2stR}	B_{x2stR}	B_{SSW}	B_M	B_M
Stocks de blé (toutes variétés)	42.3	-6.1 (12.3)	103.2 (17.6)	-94.2 (16.5)	-0.9‡ (24.8)	19.2‡ (16.5)	10.6‡ (16.7)
Superficie enssemencée en maïs	-1.8	1.1‡ (1.1)	11.7 (1.3)	10.1 (1.1)	0.3‡ (1.2)	-4.7‡ (1.9)	-5.0 (1.5)
Stocks de maïs	-6.4	-5.4‡ (1.5)	2.4 (1.6)	0.2 (1.3)	-6.5‡ (1.6)	-7.9‡ (2.4)	-9.3‡ (2.2)
Superficie cultivable	27.0	-19.6 (8.3)	-15.0 (8.3)	7.0 (3.1)	-19.6 (8.2)	-36.8 (11.0)	-12.8 (4.0)
Capacité d'entreposage du grain	-3.37	1.4‡ (3.7)	32.3 (3.7)	29.5 (2.6)	-0.1‡ (3.9)	-6.9 (3.0)	-6.8 (2.5)
Superficie enssemencée en soja	-4.4	0.8 (0.8)	13.0 (1.0)	9.9 (0.9)	-0.3 (1.0)	-2.9 (1.1)	-2.7 (1.0)
Stocks de soja	-0.01	2.8‡ (3.1)	21.3 (2.9)	5.0 (2.3)	0.2‡ (3.5)	-11.0 (3.6)	-8.8 (3.4)
Superficie totale des terres agricoles	-20.0	-24.7‡ (10.4)	-18.8‡ (12.5)	-2.6 (7.6)	-25.7‡ (10.7)	-44.5‡ (13.4)	-21.2 (5.8)
Cheptel porcin	-0.1	-2.1 (0.9)	3.4 (1.1)	-0.0‡ (1.0)	-2.2‡ (1.1)	-2.5‡ (1.3)	-1.6‡ (1.0)
Semis de blé d'hiver	-0.6	-0.5‡ (0.4)	3.8 (0.6)	1.8 (0.5)	-1.2‡ (0.6)	1.1 (0.4)	1.1 (0.4)
Nombre de postes pour lesquels B^* est incluse dans l'I.C.	6	1	1	1	8	5	3
C.V. moyen	1.01	.30	11.1	9.5	.41	10.8	.48
Valeur moyenne de la racine carré de l'EQM	13.2	22.4	25.2	12.9	14.9	10.8	91.3
Valeur moyenne du biais relatif en valeur absolue	30.8	220.0	53.4	4.9	113.1	91.3	91.3

† Erreurs types entre parenthèses.
‡ Un intervalle de confiance à 95% contient le paramètre de la pseudo-population.

(programme d'enquêtes agricoles (EA)). Ces enquêtes visent à recueillir des données sur certains produits agricoles au niveau national et au niveau des États. De 1988 à 1990, des réinterviews ont été effectuées en décembre de chacune de ces années dans six États: l'Indiana, l'Iowa, le Minnesota, le Nebraska, l'Ohio et la Pennsylvanie; ces réinterviews avaient pour but d'évaluer l'erreur systématique de mesure dans les données recueillies par interview téléphonique assistée par ordinateur (ITAO). Les techniques de réinterview utilisées à ces trois occasions ressemblent beaucoup à celles appliquées par le Census Bureau des E.-U. (voir, par exemple, Forsman et Schreiner 1991). Cependant, contrairement au programme du Census Bureau, les études du NASS visent principalement à estimer l'erreur systématique de mesure plutôt qu'à évaluer le rendement des intervieweurs.

Comme nous l'avons mentionné plus haut, seules les unités répondantes qui avaient subi une interview téléphonique assistée par ordinateur lors des EA pouvaient être sélectionnées pour faire partie de l'échantillon de réinterview. Cette restriction était surtout motivée par des raisons d'économie, de temps et de commodité. Toutefois, comme une forte proportion des EA sont réalisées par ITAO, l'information qu'on a pu recueillir sur l'erreur systématique de mesure dans les EA grâce à l'échantillon de réinterview sera précieuse pour l'ensemble du programme d'enquêtes agricoles.

Tableau 1

Taille des échantillons selon la question d'enquête

Question	Taille des échantillons selon la question d'enquête		
	x	y	μ
	U	S_1	S_2

Stock de blé (toutes variétés confondues)	108,267	8,176	1,157
Superficie ensemencée en maïs	225,269	8,211	1,157
Stocks de maïs	225,269	7,990	1,115
Superficie cultivable	278,045	8,274	1,141
Capacité d'entreposage du grain	207,460	8,126	1,104
Superficie ensemencée en soja	171,761	8,211	1,156
Stocks de soja	171,761	8,113	1,130
Superficie totale des terres agricoles	276,450	8,309	1,159
Cheptel porcin	248,571	8,247	1,142
Semis de blé d'hiver	108,267	8,211	1,150

Pour effectuer les réinterviews, le NASS comptait sur une équipe formée de surveillants des opérations sur le terrain et d'intervieweurs chevronnés. Cette équipe, qui était différente de celle affectée aux interviews téléphoniques assistées par ordinateur, a effectué des réinterviews directes auprès d'un sous-échantillon des participants aux EA en

donc réécrite B sous la forme $B = \sum_{i=1}^N (Y_i - \mu_i) / N$, où $Y_i = E(y_i | i)$. Comme Y_i est inconnu et inobservable pour tous $i \in U$, B est aussi inconnu et inobservable. Par conséquent, nous allons créer une pseudo-population semblable à U , que nous désignerons par U^* , de telle sorte que $B^* = E^*(y_i - \mu_i)$ soit connu, $E^*(\cdot)$ étant l'espérance mathématique par rapport à la distribution des erreurs de mesure et à la distribution des erreurs d'échantillonnage rattachées à U^* .

Soit $U^* = \cup_{h=1}^L U_h^*$, où U_h^* est formé de $k_h = N_h / n_{1h}$ répliques des unités de S_{1h} . Nous supposons ici que k_h est un nombre entier, mais nous allons assouplir cette hypothèse plus loin. De plus, désignons par y_i^* la valeur de la caractéristique pour l'unité $i \in U^*$. Cette valeur est égale à la valeur y_i pour l'unité correspondante dans S_1 . Par conséquent, le total des y_i^* pour la population est $Y^* = \sum_{i \in U^*} y_i^* = Y_{1st}^*$, Y_{1st}^* étant défini en (2.13). De la même manière, définissons la valeur vraie pour l'unité $i \in U^*$ comme $\mu_i^* = \mu_i$, pour $i \in U^*$ correspondant à $j \in S_2$. Pour $j \in S_{1-2}$, μ_j^* est inconnu; cependant, nous pourrions, en ce qui concerne notre pseudo-population, générer des pseudo-valeurs pour μ_j^* , de telle sorte que $M^* = \sum_{i \in U^*} \mu_i^* = \bar{M}_{2st}$, où \bar{M}_{2st} est défini en (2.11). Par conséquent, pour U^* , $B^* = Y_{1st}^* - \bar{M}_{2st}$, défini en (2.13). Comme nous le verrons, il n'est pas nécessaire de générer des pseudo-valeurs pour μ_i^* lorsqu'on veut évaluer le biais des estimateurs de B^* .

Notons que selon un plan d'échantillonnage stratifié, $U^* = U_A^*$, conformément à ce qui est mentionné dans la section 2.4. En outre, la méthode d'auto-amorçage décrite dans cette section équivaut à un échantillonnage répété dans U^* et les estimateurs $\hat{\theta}_1, \dots, \hat{\theta}_p$ de B peuvent aussi être vus comme des estimateurs de B^* . Comme B^* est connu, le biais de $\hat{\theta}$, en tant qu'estimateur de B^* , est $B^* - \hat{\theta} - B^*$ et l'EQM correspondante peut être estimée par la formule

$$EQM = \sum_q (\hat{\theta}^q - B^*)^2 / Q$$
$$= var_{BSS}(\hat{\theta}) + (\hat{\theta}^* - B^*)^2, \quad (2.35)$$

où $var_{BSS}(\hat{\theta})$, $\hat{\theta}^q$, et $\hat{\theta}^*$ sont définis dans la section 2.4. On peut vérifier facilement que ces résultats sont aussi valides lorsque k_h est un nombre non entier.

En conclusion, la méthode d'auto-amorçage offre un moyen d'évaluer l'EQM de divers estimateurs de B^* . En outre, la pseudo-population U^* est une reconstitution de U basée sur des répliques des valeurs relatives aux unités de S_1 et S_2 . Il est donc raisonnable de se servir de B^* et de l'EQM* pour évaluer divers estimateurs de B .

3. APPLICATION À L'ENQUÊTE AGRICOLE

3.1 Description de l'enquête

Le National Agricultural Statistics Service (NASS) effectue annuellement une série d'enquêtes qui est connue sous le nom de Agricultural Survey (AS) Program

puisse s'appliquer à des estimateurs de tous degrés de complexité, suivant des hypothèses qui sont cohérentes et qui ne dépendent d'aucune hypothèse de modèle. Il est notoire que les méthodes d'estimation de la variance basées sur un modèle sont très sensibles à la défaillance du modèle (voir, par exemple, Royall et Herson 1973; Royall et Cumberland 1978; et Hansen, Madow et Tepping 1983). Dans Royall et Cumberland (1981), on examine plusieurs solutions dans la perspective de l'erreur systématique, notamment l'estimateur jackknife.

Notre approche ressemble à celle de Royall et Cumberland sauf que nous utilisons un estimateur d'auto-amorçage au lieu d'un estimateur jackknife. Pour des observations indépendantes et identiquement distribuées, Efron et Gong (1983) montrent que cet estimateur diffère de l'estimateur jackknife par un facteur de $n/(n-1)$ pour des échantillons de taille n . Par conséquent, les caractéristiques de robustesse que décrivent Royall et Cumberland pour l'estimateur jackknife s'appliquent aussi à l'estimateur utilisé.

D'autres caractéristiques de l'estimateur d'auto-amorçage nous ont amenés à le préférer à d'autres méthodes de rééchantillonnage. La méthode jackknife et la méthode BRR (balanced repeated replication) sont différentes de notre étude. Au contraire, la méthode d'auto-amorçage s'adapte facilement à l'échantillonnage à deux phases. De plus, grâce à une étude de simulation, Rao et Wu (1988) ont démontré que le niveau d'efficacité des intervalles de confiance se compare avantageusement à celui des intervalles de confiance jackknife et BRR en ce qui concerne les plans d'échantillonnage complexes.

Dans le cheminement de l'étude, nous étendons la méthode élaborée par Bickel et Freedman (1984) pour l'échantillonnage à une phase stratifié à l'échantillonnage à deux phases stratifié. Puisque la méthode d'auto-amorçage est exécutée pour chaque strate prise individuellement, pour des raisons de simplicité nous allons décrire la méthode pour le cas à une strate.

2.4.1 Estimation de la variance

Étendre la méthode d'auto-amorçage à l'échantillonnage à deux phases n'équivaut pas simplement à soumettre les échantillons à une phase à un nouvel échantillonnage. Rappelons-nous que l'on connaît des valeurs vraies uniquement pour les unités de S_2 et qu'en conséquence, l'échantillonnage doit nécessairement se limiter aux unités contenues dans S_2 . Désignons donc par S_1 et S_2 les échantillons de la phase 1 et de la phase 2 respectivement, tirés de U au moyen d'un EASSR. Posons S_{1-2} comme l'ensemble d'unités, $S_1 \sim S_2$. Posons $\hat{\theta} = \hat{\theta}(S_{1-2}, S_2)$ comme un estimateur de θ qui peut être une fonction des observations relatives aux unités contenues dans S_2 aussi bien que S_{1-2} . Définissons N_1, n_2 et n_{1-2} comme les tailles respectives des ensembles U, S_1, S_2 et S_{1-2} . Voyons comment on applique la méthode d'auto-amorçage pour obtenir des estimations de $\text{Var}(\hat{\theta})$.

Le cas le plus simple est celui où N/n_1 est un nombre entier, disons k . Premièrement, nous constituons l'ensemble d'unités de la pseudo-population

La méthode d'auto-amorçage permet aussi d'estimer les biais d'un estimateur. L'estimateur habituel du biais (voir Efron et Gong 1983; Rao et Wu 1988) est $b(\hat{\theta}) = \hat{\theta}^* - \hat{\theta}$, où $\hat{\theta}^* = \sum \hat{\theta}_q^*/Q$ et $\hat{\theta}$ est l'estimation calculée à l'aide de l'échantillon complet. Notons que $\hat{\theta}_q^*(q = 1, \dots, Q)$ et $\hat{\theta}$ ont la même forme fonctionnelle et reposent sur les mêmes hypothèses de modèle. Par conséquent, $b(\hat{\theta})$ ne tient pas compte de l'effet de la défaillance du modèle sur le biais. Nous proposons donc un autre estimateur du biais qui, croyons-nous, est supérieur à $b(\hat{\theta})$.

Rappelons-nous, d'après l'équation (2.4), que $\bar{B} = E(y_i - \mu_i)$, où $E(\cdot)$ désigne l'espérance mathématique par rapport à la distribution des erreurs d'échantillonnage et la distribution des erreurs de mesure. Nous pouvons

2.4.2 Estimation du biais et de l'EOM

$$\alpha = \left(1 - \frac{n_1}{r}\right) \left(1 - \frac{N-1}{r}\right). \quad (2.34)$$

En nous servant des méthodes de Rao et Wu (1988), nous pouvons maintenant montrer que $\text{var}_{BSS}(\hat{\theta})$ est un estimateur convergent de $\text{Var}(\hat{\theta})$. Si $N = kn_1 + r$, où $0 < r < n_1$, on doit procéder d'une autre manière en utilisant la méthode de Bickel et Freedman. Premièrement, on constitue, comme précédemment, la pseudo-population U_A^* , formée de kn_1 unités. De plus, on crée la pseudo-population $U_B^* = U_{B(1-2)} \cup U_{B(2)}^*$, de taille $(k+1)n_1$, où $U_{B(1-2)}^*$ et $U_{B(2)}^*$ sont composés de $(k+1)$ répliques des unités de S_{1-2} et de S_2 respectivement. Alors, pour un nombre αQ d'échantillons obtenus par la méthode d'auto-amorçage, tirons $S_1^* = S_{1-2}^* \cup S_2^*$ dans U_A^* , et pour $(1-\alpha)Q$ échantillons, tirons S_1^* dans la pseudo-population U_B^* à l'aide de la procédure à trois étapes décrite ci-dessus, où

$$\text{var}_{BSS}(\hat{\theta}) = \sum_{q=1}^Q \frac{(\hat{\theta}_q^* - \hat{\theta}^*)^2}{Q - 1}, \quad \text{où } \hat{\theta}^* = \sum_{q=1}^Q \hat{\theta}_q^*/Q. \quad (2.33)$$

Nous répétons les étapes ci-dessus un nombre élevé de fois, Q , pour obtenir $\hat{\theta}_1^*, \dots, \hat{\theta}_Q^*$. Nous pouvons ainsi définir un estimateur de $\text{Var}(\hat{\theta})$ par la formule

1. Tirer un EASSR de taille n_2 dans $U_A^*(2)$ et désigner cet ensemble par S_2^* .
2. Tirer un EASSR de taille n_{1-2} dans $U_A^*(1-2)$ et désigner cet ensemble par S_{1-2}^* .
3. Calculer $\hat{\theta}_1^* = \hat{\theta}_1(S_{1-2}^*, S_2^*)$, qui a la même forme fonctionnelle que $\hat{\theta}(S_{1-2}, S_2)$ mais est calculé pour les $n_1 = n_{1-2} + n_2$ unités contenues dans $S_1^* = S_{1-2}^* \cup S_2^*$.

Nous répétons les étapes ci-dessus un nombre élevé de fois, Q , pour obtenir $\hat{\theta}_1^*, \dots, \hat{\theta}_Q^*$. Nous pouvons ainsi définir un estimateur de $\text{Var}(\hat{\theta})$ par la formule

$$U_A^* = U_A^*(2) \cup U_A^*(1-2), \quad (2.32)$$

qui équivaut à $B_{SSW}^{x2sr} = B_{SSW}$ plus le second terme du membre de droite de l'équation (2.21).

2.3.2 Estimateurs non pondérés de M et B

Réécrivons M sous la forme

$$M = \sum_{i \in S_2} \mu_i + \sum_{i \in S_1 \sim S_2} \mu_i + \sum_{i \in U \sim S_1} \mu_i \quad (2.23)$$

$$= M^{(2)} + M^{(1 \sim 2)} + M^{(\sim 1)},$$

par exemple, où $S_g = \bigcup_{h=1}^g S_{gh}$, $g = 1, 2$. La façon de procéder pour l'estimation non pondérée basée sur un modèle est de remplacer μ_i dans $M^{(1 \sim 2)}$ et $M^{(\sim 1)}$ par une prévision, $\hat{\mu}_i$, tirée d'un modèle.

Si nous utilisons le modèle (2.14), nous avons comme estimateur de μ_i

$$\hat{\mu}_i = y_i / \gamma_i,$$

où, désormais, $\gamma_i = \bar{y}_2 / \mu_2$. Nous avons donc comme estimateur de $M^{(1 \sim 2)}$

$$\hat{M}^{(1 \sim 2)} = \frac{\bar{\mu}_2}{\bar{y}_2} \sum_{i \in S_1 \sim S_2} y_i$$

(2.24)

$$= \frac{\bar{\mu}_2}{\bar{y}_2} (n_1 \bar{y}_1 - n_2 \bar{y}_2),$$

où $\bar{y}_g = \sum_{i \in S_g} y_i / n_g$, $\bar{\mu}_2 = \sum_{i \in S_2} \mu_i / n_2$, et $n_g = \sum_h n_{gh}$ pour $g = 1, 2$. De plus, en utilisant le modèle

$$(2.25) \quad \mu_i = \delta x_i + \xi_i,$$

où δ est une constante et $\xi_i \sim (0, \sigma_\xi^2 x_i)$, nous obtenons

$$(2.26) \quad \hat{M}^{(\sim 1)} = \frac{\bar{x}_2}{\bar{\mu}_2} X_{U \sim S_1},$$

où $X_{U \sim S_1} = \sum_{i \in U \sim S_1} X_i$. Nous avons donc comme estimateur basé sur un modèle de M

$$\hat{M}_M = M^{(2)} + \hat{M}^{(1 \sim 2)} + \hat{M}^{(\sim 1)}$$

(2.27)

$$= M^{(1)} + \hat{M}^{(\sim 1)},$$

où $\hat{M}^{(1)} = n_1 \bar{\mu}_2 \bar{y}_1 / \bar{y}_2$.

De la même manière, nous pouvons réécrire Y sous la forme

$$Y = \sum_{i \in S_1} y_i + \sum_{i \in U \sim S_1} y_i$$

(2.28)

$$= Y^{(1)} + Y^{(\sim 1)}$$

et nous cherchons à prédire y_i dans $Y^{(\sim 1)}$. En utilisant le modèle (2.18), nous avons comme estimateur basé sur un modèle de $Y^{(\sim 1)}$

$$Y^{(\sim 1)} = \frac{\bar{x}_1}{\bar{y}_1} X_{U \sim S_1}$$

et, par conséquent, comme estimateur de Y

$$(2.29) \quad Y_M = Y^{(1)} + Y^{(\sim 1)}.$$

B est donc estimé au moyen de la formule

$$(2.30) \quad B_M = Y_M - M_M.$$

On peut aussi construire des versions de B_{2sr} , B_{12sr} , B_{x2sr} et B_M qui résistent mieux aux valeurs aberrantes de modèle. On peut créer ces estimateurs, désignés par B_{2sr} , B_{12sr} , B_{x2sr} et B_M respectivement, en supprimant les points de données qui s'écartent sensiblement des prédictions du modèle et en calculant les estimateurs basés sur un modèle ou les estimateurs dépendants d'un modèle à l'aide des données restantes. À titre d'exemple, prenons l'estimateur M_{2sr} défini en (2.15). Pour cet estimateur, posons

$$(2.31) \quad \sum_{\mu_{hi} \neq 0} (y_{hi} - \hat{\gamma} \mu_{hi})^2 = (n_2 - 1) S_{res,h}^2 = \frac{\sum_{\mu_{hi} \neq 0} (\mu_{hi} - \hat{\gamma} \mu_{hi})^2}{\sum_{\mu_{hi} \neq 0} (\mu_{hi} - \hat{\gamma} \mu_{hi})^2},$$

comme la somme des carrés des résidus pour le modèle (2.14). Ensuite, pour calculer l'estimateur de γ , on n'utilise que les unités désignées par $i \in S_{2h}$, où $S_{2h} = \{i \in S_{2h} : |y_{hi} - \hat{\gamma} \mu_{hi}| \leq 3 s_{res,h} \sqrt{\mu_{hi}}\}$. Si l'on désigne cet estimateur par $\hat{\gamma}$, l'estimateur de M est $M_{2sr} = X_{1sr} / \bar{\gamma}$, où $\bar{\gamma} = \bar{y}_{2sr} / \bar{\mu}_{2sr}$ et $\bar{\mu}_{2sr}$ et \bar{y}_{2sr} sont les moyennes stratifiées de μ_i et y_i pour $i \in S_{2h}$. Les autres estimateurs robustes se calculent de la même façon.

De nombreux autres estimateurs non pondérés basés sur un modèle peuvent être examinés par rapport à l'échantillonnage à deux phases. On peut, par exemple, ajouter une ordonnée à l'origine dans les modèles (2.14), (2.18) et (2.26). De plus, on peut définir séparément des paramètres de pente et d'ordonnée à l'origine pour chaque strate ou chaque combinaison de strates.

2.4 Estimation des erreurs quadratiques moyennes à l'aide d'estimateurs d'auto-amorçage

Bien qu'il soit possible, suivant les hypothèses basées sur un plan ou basées sur un modèle appropriées, de calculer des estimations en forme analytique de la variance des estimateurs considérés dans cette étude, nous avons préféré utiliser une méthode de rééchantillonnage intensif sur ordinateur. Primo, nous cherchons une méthode qui est facile à appliquer puisque notre étude est susceptible de porter sur un grand nombre d'estimateurs. Secundo, il est essentiel d'évaluer chaque estimateur à l'aide des mêmes critères et à cette fin, une méthode cohérente d'estimation de la variance est indispensable. Nous devons donc employer une méthode d'estimation de la variance qui

sans difficultés ce raisonnement aux modèles multidimensionnels avec ordonnée à l'origine; ce type de généralisation fera l'objet d'un article futur. Donc, si nous posons $\gamma_0 = 0$ dans (2.1), nous obtenons

$$(2.14) \quad y_i = \gamma \mu_i + \epsilon_i,$$

où γ est une constante inconnue et, par hypothèse, $\epsilon_i \sim (0, \sigma_e^2 \mu_i)$. L'estimateur gaussien de γ est $\hat{\gamma} = \hat{y}_{2st} / \hat{\mu}_{2st}$, où $\hat{y}_{2st} = \bar{y}_{2st} / N = \bar{M}_{2st} / N$. Par conséquent, un estimateur dépendant d'un modèle de μ_i est $y_i / \hat{\gamma} = \hat{\mu}_{2st} y_i / \hat{y}_{2st}$ et un estimateur équivalent de M est

$$(2.15) \quad M_{2stR} = \frac{\bar{M}_{2st}}{\bar{y}_{2st}} \bar{y}_{1st}.$$

Si nous utilisons cet estimateur de M , voici deux estimateurs de B qui correspondent à (2.12) et à (2.13):

$$(2.16) \quad B_{2stR} = \bar{y}_{2st} - M_{2stR}$$

$$(2.17) \quad B_{12stR} = \bar{y}_{1st} - M_{2stR}.$$

On peut obtenir un troisième estimateur de B à l'aide du modèle

$$(2.18) \quad y_i = \beta x_i + \epsilon_i,$$

où β est une constante et $\epsilon_i \sim (0, \sigma_e^2 x_i)$. Cela donne un estimateur par quotient de y_i ,

$$(2.19) \quad \bar{y}_{xstR} = \frac{\bar{y}_{1st}}{\bar{x}_{1st}} \bar{x}_i.$$

Par conséquent, l'estimateur correspondant de B est

$$(2.20) \quad B_{x2stR} = \bar{y}_{xstR} - M_{2stR}.$$

Enfin, Särndal, Swensson et Wretman (1992, p. 360) proposent un estimateur général de M pour l'échantillonnage à deux phases. En appliquant leur équation 9.7.2 au modèle (2.14) suivant un échantillonnage stratifié, nous obtenons

$$(2.21) \quad M_{SSW} = \bar{M}_{2st} + \frac{\bar{\mu}_{2st}}{\bar{\mu}_{1st}} (X - \bar{x}_{1st}).$$

Notons que l'estimateur ci-dessus n'est ni plus ni moins non biaisé de la valeur nulle. Le nouvel estimateur peut avoir une variance moins élevée que celle de \bar{M}_{2st} si le second terme du membre de droite de l'équation (2.21) est corrélé négativement avec \bar{M}_{2st} . De la même manière, l'estimateur de X que proposent les auteurs ci-dessus se ramène à \bar{y}_{xstR} , défini en (2.19). Par conséquent, l'estimateur correspondant de B est

$$(2.22) \quad B_{SSW} = \bar{y}_{xstR} - M_{SSW},$$

phases est prélevé dans chaque strate au moyen d'un échantillonnage aléatoire simple à chaque phase. Désignons par n_{1h} et $n_{2h} \leq n_{1h}$ la taille des échantillons des phases 1 et 2 respectivement dans la strate h . Soient S_{1h} et $S_{2h} \subseteq S_{1h}$ les ensembles d'éléments pour les échantillons des phases 1 et 2 respectivement dans la strate h . Supposons que les données suivantes ont été observées ou qu'elles sont connues d'une autre manière:

variables de résultat: $y_i \quad \forall i \in S_{1h}$

valeurs vraies: $\mu_i \quad \forall i \in S_{2h}$

variables auxiliaires: $x_i \quad \forall i \in S_{1h}$.

Supposons en outre que $X_h = \sum_{i \in U_h} x_i$ est connu, pour $h = 1, \dots, L$, où U_h est l'ensemble d'éléments pour la strate h .

2.3.1 Estimateurs pondérés de M et B

Pour des raisons de commodité, nous allons considérer l'estimation du biais de l'estimateur d'un total de population désigné par M . L'estimateur habituel de $M = NM$ est l'estimateur stratifié sans biais défini par l'équation

$$(2.11) \quad M_{2st} = \sum_h N_h \bar{\mu}_{2h},$$

où $\bar{\mu}_{2h} = \sum_{i \in S_{2h}} \mu_i / n_{2h}$. L'estimateur correspondant de $B = NB$ équivalant à N fois le TEN défini en (1.1). Pour les échantillons stratifiés, l'estimateur est

$$(2.12) \quad B_{2st} = \bar{y}_{2st} - M_{2st},$$

où $\bar{y}_{2st} = \sum_h N_h \bar{y}_{2h}$ et $\bar{y}_{2h} = \sum_{i \in S_{2h}} y_i / n_{2h}$. Notons que (2.12) ne renferme pas l'information sur y pour les unités désignées par $i \in S_{1h} \sim S_{2h}$. En revanche, l'estimateur ci-dessous utilise toute l'information sur y :

$$(2.13) \quad B_{12st} = \bar{y}_{1st} - M_{2st},$$

$$\text{où } \bar{y}_{1st} = \sum_h N_h \bar{y}_{1h} \text{ et } \bar{y}_{1h} = \sum_{i \in S_{1h}} y_i / n_{1h}.$$

On peut définir un certain nombre d'estimateurs dépendants d'un modèle pour des plans à deux phases stratifiés. Ces estimateurs peuvent être de deux types: indépendants ou combinés (voir, par exemple, Cochran 1977, pp. 327-330). En outre, on peut appliquer les coefficients de correction aux estimateurs de strate de la phase 1 ou de la phase 2. Comme les tailles d'échantillons de strate sont généralement faibles dans les échantillons à deux phases, nous n'étudierons ici que des estimateurs combinés.

Etant donné que cet article insiste sur l'élaboration de méthodes d'estimation de l'erreur systématique de mesure par la prédiction modeliste et sur leur évaluation, nous allons considérer un cas simple, notamment un cas partiellement du modèle (2.1), où $\gamma_0 = 0$ (c.-à-d. le modèle "sans ordonnée à l'origine"). Néanmoins, on peut étendre

satisfait, alors le TFEN est un estimateur non biaisé de B . On peut montrer en outre que la variance du TFEN est

$$E \left\{ \left(1 - \frac{n_1}{n_2} \right) \frac{s_y^2}{s_x^2} \left(1 - \frac{s_y^2}{s_x^2} \right) \right\} + \left(1 - \frac{n_1}{n_2} \right) \frac{s_y^2}{s_x^2} (1 - r)^2, \quad (2.6)$$

où $s_x^2 = \sum_{j \in S_2} (\mu_j - \bar{\mu}_2)^2 / (n_2 - 1)$, avec des définitions analogues pour s_y^2 et s_{xy} , et $r = s_{xy} / s_y^2$.

Le TFEN peut être sous-optimal dans un certain nombre de situations qui se reproduisent périodiquement. Pour nous en convaincre, considérons des estimateurs de B de la forme suivante

$$\bar{b}_{ga} = y_g - \bar{\mu}_{Ra}, \quad (2.7)$$

où $y_g = \sum_{j \in S_g} y_j / n_g$, $g = 1, 2$,

$$\bar{\mu}_{Ra} = \bar{\mu}_2 + a(y_1 - y_2) \quad (2.8)$$

et $\bar{\mu}_2 = \sum_{j \in S_2} \mu_j / n_2$, pour a , une constante, étant donné le sous-échantillon S_1 . On peut montrer que la valeur de a qui minimise $\text{Var}(\bar{b}_{ga})$ est

$$a = r \quad \text{pour } g = 1, \\ \text{ou} \quad a = r - 1 \quad \text{pour } g = 2. \quad (2.9)$$

Par conséquent, pour $g = 1$ ou 2 , la valeur "optimale" de \bar{b}_{ga} est

$$\bar{b}^{\text{opt}} = y_1 - [\bar{\mu}_2 + r(y_1 - y_2)], \quad (2.10)$$

qui diffère de TFEN par le terme $(r - 1)(y_1 - y_2)$. Comme, en règle générale, $y_1 \neq y_2$, TFEN est optimal seulement si $r = 1$. On peut montrer que ce cas correspond à celui où γ_1 est égal à 1 dans l'équation (2.1).

Dans cet article, nous allons étudier d'autres estimateurs que TFEN qui renferment de l'information sur y pour les unités de l'ensemble $S_1 \sim S_2$ ainsi que de l'information sur des variables auxiliaires x . Afin d'illustrer les concepts, nous allons nous limiter au départ aux modèles linéaires "sans ordonnée à l'origine", c'est-à-dire aux modèles pour lesquels $\gamma_0 = 0$ dans l'équation (2.1). Cette importante classe de modèles comprend l'estimateur par différence de même que les estimateurs par quotient.

2.2 Méthodes d'estimation fondées sur la prédiction modeliste

La littérature statistique traite abondamment, en ce qui regarde le sondage de populations finies, les méthodes d'estimation de paramètres de population fondées sur la prédiction modeliste. Cochran (1977) et d'autres auteurs ont démontré les fondements de modèle de l'estimateur

par quotient, d'usage très courant. Il existe aussi beaucoup d'ouvrages dans lesquels on pose le choix entre l'utilisation de poids qui découlent d'hypothèses explicites de modèle dans l'estimation pour enquêtes complexes et la suppression des poids d'échantillonnage. Les adeptes de l'estimation dite "à modèles" déconseillent l'utilisation de poids dans l'estimation de paramètres (voir, par exemple, Royall et Herson 1973 et Royall et Cumberland 1981). Ils prétendent qu'en ce qui concerne le sondage de populations finies, les probabilités d'échantillonnage — qu'elles soient égales ou inégales — importent peu une fois que l'échantillon est formé. Les critères de fiabilité utilisés pour les échantillons fondés sur un modèle découlent des hypothèses de distribution du modèle plutôt que des distributions d'échantillonnage. Si on choisit un modèle adéquat pour décrire la relation entre la variable de réponse et d'autres variables étudiées dans l'enquête, on peut obtenir des estimateurs "non biaisés selon le modèle" de paramètres de la population qui sont plus fiables que des estimateurs comprenant des poids.

En contrepartie, il y a les adeptes de l'échantillonnage fondés sur un plan. Au lieu de reconnaître les hypothèses fondées sur un modèle, ce groupe de personnes suppose qu'un estimateur tiré d'une enquête compte comme une réalisation parmi une vaste population de réalisations possibles et que chaque réalisation dépend de l'échantillon formé. La distribution des valeurs de l'estimateur est appelée "distribution d'échantillonnage de l'estimateur" lorsqu'on tient compte de tous les échantillons qui peuvent être constitués à l'aide du plan d'échantillonnage. Les critères qui servent à évaluer les estimateurs selon cette approche portent alors sur les caractéristiques des distributions d'échantillonnage des estimateurs. Dans ce contexte, la pondération des estimateurs est indispensable pour obtenir des estimateurs sans biais, si on a recours à l'échantillonnage avec probabilités inégales.

Bien que les estimateurs de B considérés ici représentent les trois catégories d'estimateurs, le but de cet article n'est pas nécessairement de comparer les estimateurs de chaque catégorie, c'est-à-dire estimateurs fondés sur un plan, estimateurs dépendants d'un modèle et estimateurs fondés sur un modèle. Nous cherchons plutôt, en premier lieu, à élaborer une méthode systématique pour évaluer des estimateurs pour un plan d'échantillonnage à deux phases donné. Nous posons le problème en ces termes: étant donné un plan d'échantillonnage à deux phases et des estimateurs de $B = NB$ désignés par B_1, B_2, \dots, B_p , comment l'analyste détermine-t-il l'estimateur à erreur quadratique moyenne minimum? Comme deuxième objectif, nous nous proposons de définir un certain nombre d'estimateurs et d'en faire l'évaluation à l'aide d'une méthode systématique. À titre d'exemple, nous servons de données de l'enquête agricole réalisée en décembre 1990 par le National Agricultural Statistics Service.

2.3 Estimateurs considérés dans l'étude

Si nous étendons la notation définie plus haut aux plans à deux phases stratifiés, désignons par N_h la taille de la strate h , pour $h = 1, \dots, L$. Un échantillon à deux

2. MÉTHODES D'ESTIMATION ET D'ÉVALUATION

2.1 Modèle de l'erreur de mesure

Plus précisément, nous allons étudier le cas de l'échantillonnage aléatoire simple sans remise (EASSR) dans une population unique. Plus loin, nous étendrons notre raisonnement au cas de l'échantillonnage aléatoire stratifié, ce qui est assez simple.

Désignons par $U = \{1, 2, \dots, N\}$ l'ensemble des indices pour la population et désignons par $S_1 = \{1, 2, \dots, n_1\}$ sans perte de généralité, l'ensemble des indices pour l'EASSR de première phase formé de n_1 unités de U .

Pour $y_i, i \in S_1$, supposons le modèle

$$(2.1) \quad y_i = \gamma_0 + \gamma\mu_i + \epsilon_i,$$

où μ_i est la valeur vraie de la caractéristique étudiée, γ_0 et γ sont des constantes et ϵ_i est un terme d'erreur indépendant d'espérance nulle et de variance conditionnelle $\sigma_{\epsilon_i}^2$. Puisque notre analyse porte précisément sur le biais rattaché aux valeurs y_i , considérons l'espérance de y_i . Soit $E(y_i | i)$ l'espérance conditionnelle de y_i pour la distribution des ϵ_i , l'unité i étant fixe, et soit $E(y_i) = E[E(y_i | i)]$ l'espérance de $E(y_i | i)$ pour la distribution d'échantillonnage. Alors, pour une unité donnée i ,

$$(2.2) \quad E(y_i | i) = \gamma_0 + \gamma\mu_i$$

et, par conséquent, l'espérance inconditionnelle est

$$(2.3) \quad E(y_i) = \gamma_0 + \gamma\bar{M},$$

où $\bar{M} = \sum_{i=1}^N \mu_i / N$. L'erreur systématique de mesure s'écrit donc

$$(2.4) \quad \bar{B} = E(y_i - \mu_i) = \gamma_0 + (\gamma - 1)\bar{M}.$$

Le paramètre γ_0 est un terme d'erreur systématique constant qui ne dépend pas de la grandeur de \bar{M} . Notons que cette définition de γ_0 est conforme à la définition de l'erreur systématique de mesure qui découle normalement du modèle élémentaire

$$(2.5) \quad y_i = \mu_i + \epsilon_i,$$

avec $\epsilon_i \sim (\gamma_0, \sigma_{\epsilon_i}^2)$. (Voir, par exemple, Biemer et Stokes 1991.)

Considérons l'estimation de \bar{B} . Supposons qu'un sous-échantillon de taille n_2 est prélevé parmi les n_1 unités de l'échantillon initial et que la valeur vraie μ_i est déterminée pour ces n_2 unités. On peut établir la valeur vraie soit par une réinterview, une vérification d'enregistrements, une observation faite par l'intervieweur ou un autre moyen. Désignons par $S_2 \subseteq S_1$ cet échantillon de seconde phase. L'estimateur habituel de l'erreur systématique de mesure est le TEN défini en (1.1). Si l'hypothèse selon laquelle "la valeur vraie, μ_i , est observée dans la phase 2 pour tous $i \in S_2$ " est

$$(1.1) \quad \text{TEN} = \bar{y}_2 - \bar{\mu}_2,$$

de l'échantillon d'évaluation au moyen de la formule est le taux d'écart net, qui est calculé pour les n_2 répondants d'évaluation. L'estimateur habituel du biais dans les réponses d'unités qui constituent le sous-échantillon ou l'échantillon participants à la première enquête et par n_2 le nombre de caractéristiques à l'étude. Désignons par n_1 le nombre de déterminé par un moyen quelconque la valeur vraie des échantillon aléatoire des participants à l'enquête et on dans une étude de réévaluation typique, on tire un sous-notre approche peut être décrit dans les termes suivants.

Dans cet article, nous proposons de considérer des estimateurs du biais dans les réponses qui sont plus efficaces que les estimateurs classiques. Le fondement de pas l'être pour l'estimation de l'erreur rattachée à de faibles sous-populations ou à des caractéristiques peu courantes.

Comme nous avons ici un plan d'échantillonnage à deux phases stratifié et les données correspondantes (y, μ, x) , notre objectif est de déterminer le "meilleur" estimateur de l'erreur systématique de mesure, étant donné ces observations. Il s'agit essentiellement de définir un modèle pour la valeur vraie, μ_i , qui soit une fonction des valeurs observées $y_i, i = 1, \dots, n_1$, et de toute information supplémentaire, x , qui peut exister sur la population. Ce modèle sert ensuite à prédire μ_i pour toutes les unités de la population pour lesquelles cette valeur est inconnue. On peut par la suite utiliser ces prévisions pour calculer des estimations de la moyenne, du total ou de la proportion de population vrais. On peut donc déterminer des estimateurs du biais dans les réponses pour ces paramètres à partir de l'enquête principale. Comme la méthode produit une équation prédictive pour μ_i qui est une fonction des observations, on peut calculer des estimateurs du biais dans les réponses pour des domaines à taille d'échantillon faible. Dans de telles circonstances, on peut ajouter des variables géographiques et des variables individuelles dans l'équation prédictive pour μ_i (par ex., caractéristiques démographiques, type d'unité, taille de l'unité, caractéristiques géographiques, et ainsi de suite).

Dans la section qui suit, nous exposons les principes d'estimation et d'évaluation qui sous-tendent la méthode prédictive appliquée à l'estimation du biais dans les réponses. Suivant un plan d'échantillonnage aléatoire stratifié, nous produisons des estimateurs de moyennes et de totaux ainsi que les variances et les erreurs quadratiques moyennes correspondantes. Nous présentons aussi les résultats de l'application de la méthode à des données du National Agricultural Statistics Service (NAASS).

Estimation de l'erreur systématique de mesure par la prédiction modeliste

PAUL P. BIEMER et DALE ATKINSON¹

RÉSUMÉ

Les méthodes qui servent à estimer le biais de réponse dans les enquêtes requièrent des mesures répétées "non biaisées" pour à tout le moins un sous-échantillon d'observations. L'estimateur habituel du biais de réponse est la différence entre la moyenne des observations originales et la moyenne des observations non biaisées. Dans cet article, nous étudions divers estimateurs du biais de réponse tirés de la prédiction modeliste. Nous supposons comme plan de sondage un échantillonnage à deux phases stratifié, avec échantillonnage aléatoire simple dans chaque phase. Nous supposons que la caractéristique y est observée pour chaque unité échantillonnée dans la phase 1, tandis que la valeur vraie de la caractéristique, μ , est observée pour chaque unité du sous-échantillon prélevée dans la phase 2. Nous supposons en outre qu'une variable auxiliaire x est connue pour chaque unité de l'échantillon de la phase 1 et que le chiffre de population de x est connu. On suppose un certain nombre de modèles de biais de réponse, μ , et de ces modèles découlent divers estimateurs de $E(y - \mu)$, le biais de réponse. Les estimateurs sont calculés à l'aide d'une méthode d'auto-amorçage destinée à l'estimation de la variance, du biais et de l'erreur quadratique moyenne. La méthode que nous utilisons est en fait la méthode de Bickel-Freedman à une phase, étendue à un plan à deux phases stratifié. À des fins d'illustration, nous appliquons la méthode étudiée à des données du programme de réinterview du National Agricultural Statistics Service. Nous montrons par ces données que l'estimateur fondé sur un modèle de Särndal, Swensson et Wretling (1991) est supérieur à l'estimateur de différence habituel, ce qui prouve qu'il est possible d'améliorer les estimateurs classiques au moyen de la prédiction modeliste.

MOTS CLÉS: Réinterview; mesures répétées; erreur de réponse; auto-amorçage (bootstrap).

1. INTRODUCTION

Dans les ouvrages sur les enquêtes statistiques, on reconnaît que la valeur observée des caractéristiques étudiées peut différer sensiblement de la valeur réelle de ces caractéristiques lorsque l'enquête par sondage fait intervenir des répondants. Il existe un grand nombre de cas où des erreurs de mesure sont observées dans des enquêtes. Par exemple, on peut vérifier l'exactitude d'une réponse fournie en comparant cette donnée à la valeur réelle (ou, du moins, à une valeur plus exacte) de la caractéristique étudiée, contenue dans des dossiers administratifs ou des documents juridiques. Une autre méthode de vérification consiste à réinterviewer des répondants et à produire des rapports révisés. Dans une réinterview, on communique à nouveau avec un répondant pour lui faire subir une seconde interview qui vise à mesurer les mêmes caractéristiques que dans la première interview. Au lieu de reprendre simplement les questions de l'interview initiale, on peut approfondir certains questions dans le but d'obtenir des réponses plus exactes, ou on peut demander au répondant de consulter un document dans lequel figurent des valeurs possibles pour les caractéristiques étudiées. Dans le cas de certaines enquêtes de réinterview, l'intervieweur tente de résoudre, avec l'aide du répondant, les différences de

réponse entre la première et la seconde interview jusqu'à ce qu'il n'y ait plus de doute sur l'exactitude de la réponse. Forsman et Schreiner (1991) donnent un aperçu des ouvrages qui portent sur ce type de réinterviews. Parmi les autres moyens de vérifier l'exactitude des réponses d'enquête, notons: a) la comparaison des statistiques d'enquêtes (c.-à-d., moyennes, totaux, proportions, etc.) avec des statistiques provenant de sources externes, plus précises; b) l'utilisation de plans d'expérience dans le but d'estimer l'incidence des intervieweurs et des autres membres du personnel affectés aux enquêtes sur les estimations; et c) la vérification de la cohérence interne des résultats d'une enquête. Les recherches actuelles portent plus spécialement sur les estimateurs de l'erreur systématique de mesure, évaluée par rapport à des données recueillies dans des études de réévaluation – par ex., études de vérification des enregistrements et réinterviews – qui ont pour objectif de faire connaître la valeur vraie de la caractéristique à un coût (par observation) probablement beaucoup plus élevé que dans l'enquête initiale. À cause du coût normalement élevé des réinterviews, on recueille à nouveau des données uniquement pour une petite partie de l'échantillon initial. Tandis que la taille de l'échantillon peut être suffisante pour estimer l'erreur systématique au niveau national ou régional, elle peut ne

¹ Paul P. Biemer, chercheur principal, Center for Survey Research, Research Triangle Institute, Research Triangle Park (NC), E.-U., 27709; Dale Atkinson, statisticien superviseur, National Agricultural Statistics Service, 3251 Old Lee Highway, salle 305, Fairfax (VA), E.-U., 22030.

Dans ce numéro

Ce numéro de *Techniques d'enquête* contient des articles portant sur différents sujets. Dans le premier article, Biemer et Atkinson présentent une méthode générale pour construire et évaluer des estimateurs obtenus à l'aide de la prédiction modeliste de l'erreur systématique de mesure pour un plan de sondage à deux phases stratifié avec échantillonnage aléatoire simple dans chaque phase. À des fins d'évaluation, ils ont étendu la méthode bootstrap de Bickel et Freedman à l'échantillonnage à deux phases. L'exemple utilisé pour illustrer la méthode proposée montre qu'il est possible d'améliorer l'estimateur classique de la différence nette et de réaliser ainsi des économies au niveau du coût des enquêtes de réinterview. À l'origine, l'article de Armstrong et Mayda devait paraître dans la section spéciale sur le *couplage d'enregistrements et l'appariement statistique*. Les auteurs examinent l'estimation fondée sur un modèle des taux d'erreur de classification liés au couplage d'enregistrements. La classe de modèles étudiés permet la non-indépendance du statut d'appariement de différents champs d'appariement dans une paire d'enregistrements. On élabore des méthodes d'estimation et l'on en compare certaines qui portent sur l'estimation du taux d'erreur à l'aide de données synthétiques et réelles.

Pfeffermann et Bleuer étudient l'estimation pour de petites régions à l'aide de données provenant d'une enquête par panel avec renouvellement dans le temps. Leur méthode est fondée sur un modèle d'espace d'états pour les valeurs de la population dans le temps et de modèles autorégressifs distincts pour les séries d'erreurs de l'enquête provenant de chaque panel. Afin d'obtenir une certaine robustesse, on impose aux estimateurs pour de petites régions la contrainte additionnelle que leur somme doit être égale à celle des estimateurs directs de l'enquête dans de plus grandes régions définies à l'avance. La méthode est présentée à l'aide de données de l'Enquête (canadienne) sur la population active pour les provinces de l'Atlantique.

Mian et Laniel traitent de deux procédures itératives pour trouver les estimations d'un modèle d'évaluation non linéaire par la méthode du maximum de vraisemblance qui semble approprié pour des séries chronologiques économiques provenant de grosses enquêtes par sondage. Des expressions en forme analytique fermée pour les variances et les covariances asymptotiques des séries étalonées et des valeurs ajustées sont aussi fournies. Les méthodes sont illustrées à l'aide de données sur le commerce de détail au Canada.

Deville utilise des modèles de superpopulation pour anticiper, avant l'enquête, les variances d'estimation de ratios. À l'aide de modèles simples et réalistes, il produit des expressions plus ou moins complexes qu'il parvient à optimiser. Il traite du problème de l'estimation de la fréquence des erreurs dans l'ensemble des formules recueillies lors du contrôle de la qualité du recensement français.

Casady et Valliant utilisent des techniques asymptotiques pour étudier la stratification a posteriori à partir d'un point de vue conditionnel fondé sur un plan d'enquête. Les auteurs calculent, pour de grands échantillons, le biais et les erreurs quadratiques moyennes de l'estimateur de Horvitz-Thompson, d'un estimateur par quotient et d'un nouvel estimateur courant, de l'estimateur de Horvitz-Thompson. La théorie élaborée fait l'objet de tests empiriques portant sur des populations réelles et sur une population artificielle. Le problème du biais imputable aux bases de sondage défectueuses est aussi étudié.

Bandyopadhyay et Adhikari examinent l'estimation fondée sur des bases de sondage où certaines unités apparaissent plus d'une fois, chaque fois avec une identité différente. On compare les erreurs quadratiques moyennes des estimateurs fondés sur des bases parfaite et imparfaite. On étudie l'estimation d'un ratio, sur la base de sondage.

Roesch, Green et Scott présentent un concept généralisé pour toutes les méthodes couramment utilisées d'échantillonnage des forêts. Selon ce concept, la forêt est perçue comme une image bidimensionnelle découpée en pièces comme un casse-tête, les pièces étant définies par les probabilités de sélection individuelles des arbres de la forêt.

L'article de Kalton et Citro est une version révisée du discours-programme prononcé dans le cadre du Symposium '92 de Statistique Canada portant sur les enquêtes longitudinales. L'article examine comment différents plans d'enquête visant à recueillir des données dans le temps permettent d'atteindre divers objectifs analytiques. L'auteur se concentre ensuite sur les enquêtes par panel et traite des décisions qui doivent être prises au moment de la conception de ces enquêtes.

TECHNIQUES D'ENQUÊTE

Une revue de Statistique Canada
Volume 19, numéro 2, décembre 1993

TABLE DES MATIÈRES

Dans ce numéro	135
P.P. BIEMER et D. ATKINSON Estimation de l'erreur systématique de mesure par la prédiction modeliste	137
J.B. ARMSTRONG et J.E. MAYDA Estimation modeliste des taux d'erreur liés au couplage d'enregistrements	147
D. PFEFFERMANN et S.R. BLEUER Modélisation conjointe robuste de séries de données sur l'activité pour de petites régions	159
I.U.H. MIAN et N. LANIEL Estimation d'un modèle d'étalement à biais multiplicatif constant par la méthode du maximum de vraisemblance et application	175
J.-C. DEVILLE Plans de sondage à deux degrés optimaux pour des estimateurs de ratios: application au contrôle de qualité du recensement français de 1990	183
R.J. CASADY et R. VALLANT Propriétés conditionnelles des estimateurs de stratification a posteriori selon la théorie normale	193
S. BANDYOPADHYAY et A.K. ADHIKARI Échantillonnage dans des bases imparfaites contenant un nombre inconnu d'enregistrements répétés	205
F.A. ROESCH JR., E.J. GREEN et C.T. SCOTT Un nouveau concept pour l'échantillonnage des forêts	211
G. KALTON et C.F. CITRO Enquêtes par panel: ajout d'une quatrième dimension	217
Remerciements	229

TECHNIQUES D'ENQUÊTE

Une revue de Statistique Canada

La revue Techniques d'enquête est répertoriée dans The Survey Statistician et Journal Contents in Qualitative Methods. On peut en trouver les références dans Current Index to Statistics, et Journal Contents in Qualitative Methods.

COMITÉ DE DIRECTION

Président

G.J. Brackstone

Membres

B.N. Chinnappa

G.J.C. Hole

F. Mayda (Directeur de la production)

M.P. Singh

R. Platek (Ancien président)

COMITÉ DE RÉDACTION

Rédacteur en chef

M.P. Singh, *Statistique Canada*

Rédacteurs associés

D.R. Bellhouse, *University of Western Ontario*

D. Binder, *Statistique Canada*

M. Colledge, *Statistique Canada*

E.B. Dagum, *Statistique Canada*

J.-C. Deville, *INSEE*

D. Drew, *Statistique Canada*

R.E. Fay, *U.S. Bureau of the Census*

W.A. Fuller, *Iowa State University*

J.F. Gentleman, *Statistique Canada*

M. Gonzalez, *U.S. Office of Management and Budget*

R.M. Groves, *U.S. Bureau of the Census*

D. Holt, *University of Southampton*

G. Kalton, *University of Michigan*

Rédacteurs adjoints

N. Laniel, P. Lavallée, L. Mach et H. Mantel, *Statistique Canada*

POLITIQUE DE RÉDACTION

La revue Techniques d'enquête publie des articles sur les divers aspects des méthodes statistiques qui intéressent un organisme statistique comme, par exemple, les problèmes de conception découlant de contraintes d'ordre pratique, l'utilisation de différentes sources de données et de méthodes de collecte, les erreurs dans les enquêtes, l'évaluation des enquêtes, la recherche sur les méthodes d'enquête, l'analyse des séries chronologiques, la désaisonnalisation, les études démographiques, l'intégration de données statistiques, les méthodes d'estimation et d'analyse de données et le développement de systèmes généralisés. Une importance particulière est accordée à l'élaboration et à l'évaluation de méthodes qui ont été utilisées pour la collecte de données ou appliquées à des données réelles. Tous les articles seront soumis à une critique, mais les auteurs demeurent responsables du contenu de leur texte et les opinions émises dans la revue ne sont pas nécessairement celles du comité de rédaction ni de Statistique Canada.

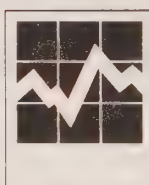
Présentation de textes pour la revue

La revue Techniques d'enquête est publiée deux fois l'an. Les auteurs désirant faire paraître un article sont invités à en faire parvenir le texte au rédacteur en chef, M. M.P. Singh, Division des méthodes d'enquêtes sociales, Statistique Canada, Tunney's Pasture, Ottawa (Ontario), Canada K1A 0T6. Prière d'envoyer quatre exemplaires dactylographiés selon les directives présentées dans la revue. Ces exemplaires ne seront pas retournés à l'auteur.

Abonnement

Le prix de la revue Techniques d'enquête (catalogue n° 12-001) est de 45 \$ par année au Canada, 50 \$ (É.-U.) aux États-Unis, et de 55 \$ (É.-U.) par année à l'étranger. Prière de faire parvenir votre demande d'abonnement à Section des ventes des publications, Statistique Canada, Ottawa (Ontario), Canada K1A 0T6. Un prix réduit est offert aux membres de l'American Statistical Association, l'Association Internationale de Statisticiens d'Enquête et la Société Statistique du Canada.

TECHNIQUES D'ENQUÊTE



UNE REVUE ÉDITÉE PAR STATISTIQUE CANADA

DÉCEMBRE 1993 • VOLUME 19 • NUMÉRO 2

Publication autorisée par le ministre
responsable de Statistique Canada

© Ministre de l'Industrie, des Sciences
et de la Technologie, 1993

Tous droits réservés. Il est interdit de reproduire ou de transmettre
le contenu de la présente publication, sous quelque forme ou
par quelque moyen que ce soit, enregistré ou non, sur support
magnétique, reproduction électronique, mécanique, photographique,
ou autre, ou de l'emmagasiner dans un système de recouvrement,
sans l'autorisation écrite préalable des Services de concession
des droits de licence, Division de la commercialisation,
Statistique Canada, Ottawa, Ontario, Canada K1A 0T6.

Décembre 1993

Prix : Canada : 45 \$

États-Unis : 50 \$ US

Autres pays : 55 \$ US

N° 12-001 au catalogue

ISSN 0714-0045

Ottawa



NUMÉRO 2

•

VOLUME 19

•

DÉCEMBRE 1993

UNE REVUE
ÉDITÉE
PAR STATISTIQUE CANADA

Catalogue 12-001

TECHNIQUES D'ENQUÊTE



